Massachusetts Institute of Technology
Department of Economics
Working Paper Series

POST-ℓ_1-PENALIZED ESTIMATORS IN HIGH-DIMENSIONAL LINEAR REGRESSION MODELS

Alexandre Belloni
Victor Chernozhukov

Working Paper 10-5
January 4, 2009

Room E52-251
50 Memorial Drive
Cambridge, MA 02142

Massachusetts Institute of Technology
Department of Economics
Working Paper Series

POST-ℓ_1-PENALIZED ESTIMATORS IN HIGH-DIMENSIONAL LINEAR REGRESSION MODELS

Alexandre Belloni
Victor Chernozhukov

Working Paper 10-5
January 4, 2009

Room E52-251
50 Memorial Drive
Cambridge, MA 02142

arXiv:1001.0188v2 [math.ST] 26 Mar 2010

POST-ℓ_1-PENALIZED ESTIMATORS IN HIGH-DIMENSIONAL LINEAR REGRESSION MODELS

ALEXANDRE BELLONI AND VICTOR CHERNOZHUKOV

ABSTRACT. In this paper we study post-penalized estimators which apply ordinary, unpenalized linear regression to the model selected by first-step penalized estimators, typically LASSO. It is well known that LASSO can estimate the regression function at nearly the oracle rate, and is thus hard to improve upon. We show that post-LASSO performs at least as well as LASSO in terms of the rate of convergence, and has the advantage of a smaller bias. Remarkably, this performance occurs even if the LASSO-based model selection "fails" in the sense of missing some components of the "true" regression model. By the "true" model we mean here the best s-dimensional approximation to the regression function chosen by the oracle. Furthermore, post-LASSO can perform strictly better than LASSO, in the sense of a strictly faster rate of convergence, if the LASSO-based model selection correctly includes all components of the "true" model as a subset and also achieves a sufficient sparsity. In the extreme case, when LASSO perfectly selects the "true" model, the post-LASSO estimator becomes the oracle estimator. An important ingredient in our analysis is a new sparsity bound on the dimension of the model selected by LASSO which guarantees that this dimension is at most of the same order as the dimension of the "true" model. Our rate results are non-asymptotic and hold in both parametric and nonparametric models. Moreover, our analysis is not limited to the LASSO estimator in the first step, but also applies to other estimators, for example, the trimmed LASSO, Dantzig selector, or any other estimator with good rates and good sparsity. Our analysis covers both traditional trimming and a new practical, completely data-driven trimming scheme that induces maximal sparsity subject to maintaining a certain goodness-of-fit. The latter scheme has theoretical guarantees similar to those of LASSO or post-LASSO, but it dominates these procedures as well as traditional trimming in a wide variety of experiments.

FIRST ARXIV VERSION: December 2009.
KEY WORDS. LASSO, POST-LASSO, POST-MODEL-SELECTION ESTIMATORS.
AMS CODES. PRIMARY 62H12, 62J99; SECONDARY 62J07.

Date: First Version: January 4, 2009. Current Revision: March 29, 2010. We thank Don Andrews, Whitney Newey, and Alexandre Tsybakov as well as participants of the Cowles Foundation Lecture at the 2009 Summer Econometric Society meeting and the joint Harvard-MIT seminar for useful comments. We also would like to thank Denis Chetverikov, Brigham Fradsen, and Joonhwan Lee for thorough proof-reading and many useful suggestions on several versions of this paper.

1. Introduction

In this work we study post-model selected estimators for linear regression in high-dimensional sparse models (HDSMs). In such models, the overall number of regressors p is very large, possibly much larger than the sample size n. However, the number s of significant regressors – those having a non-zero impact on the response variable – is smaller than the sample size, that is, $s = o(n)$. HDSMs ([6], [13]) have emerged to deal with many new applications arising in biometrics, signal processing, machine learning, econometrics, and other areas of data analysis where high-dimensional data sets have become widely available.

Several papers have begun to investigate estimation of HDSMs, primarily focusing on penalized mean regression, with the ℓ_1-norm acting as a penalty function [2, 6, 10, 13, 17, 20, 19]. [2, 6, 10, 13, 20, 19] demonstrated the fundamental result that ℓ_1-penalized least squares estimators achieve the rate $\sqrt{s/n}\sqrt{\log p}$, which is very close to the oracle rate $\sqrt{s/n}$ achievable when the true model is known. [17] demonstrated a similar fundamental result on the excess forecasting error loss under both quadratic and non-quadratic loss functions. Thus the estimator can be consistent and can have excellent forecasting performance even under very rapid, nearly exponential growth of the total number of regressors p. [1] investigated the ℓ_1-penalized quantile regression process, obtaining similar results. See [9, 2, 3, 4, 5, 11, 12, 15] for many other interesting developments and a detailed review of the existing literature.

In this paper we derive theoretical properties of post-penalized estimators which apply ordinary, unpenalized linear least squares regression to the model selected by first-step penalized estimators, typically LASSO. It is well known that LASSO can estimate the mean regression function at nearly the oracle rate, and hence is hard to improve upon. We show that post-LASSO can perform at least as well as LASSO in terms of the rate of convergence, and has the advantage of a smaller bias. This nice performance occurs even if the LASSO-based model selection "fails" in the sense of missing some components of the "true" regression model. Here by the "true" model we mean the best s-dimensional approximation to the regression function chosen by the oracle. The intuition for this result is that LASSO-based model selection omits only those components with relatively small coefficients. Furthermore, post-LASSO can perform strictly better than LASSO, in the sense of a strictly faster rate of convergence, if the LASSO-based model correctly includes all components of the "true" model as a subset and is sufficiently sparse. Of course, in the extreme case, when LASSO perfectly selects the "true" model, the post-LASSO estimator becomes the oracle estimator.

Importantly, our rate analysis is not limited to the LASSO estimator in the first step, but applies to a wide variety of other first-step estimators, including, for example, trimmed LASSO, the Dantzig selector, and their various modifications. We give generic rate results that cover any first-step estimator for which a rate and a sparsity bound are available. We also give a generic result on trimmed first-step estimators, where trimming can be performed by a traditional hard-thresholding scheme or by a new trimming scheme we introduce in the paper. Our new trimming scheme induces maximal sparsity subject to maintaining a certain goodness-of-fit (goof) in the sample, and is completely data-driven. We show that our post-goof-trimmed estimator performs at least as well as the first-step estimator; for example, the post-goof-trimmed LASSO performs at least as well as LASSO, but can be strictly better under good model selection properties. It should also be noted that traditional trimming schemes do not in general have such nice theoretical guarantees, even in simple diagonal models.

Finally, we conduct a series of computational experiments and find that the results confirm our theoretical findings. In particular, we find that the post-goof-trimmed LASSO and post-LASSO emerge clearly as the best and second best, both substantially outperforming LASSO and the post-traditional-trimmed LASSO estimators.

To the best of our knowledge, our paper is the first to establish the aforementioned rate results on post-LASSO and the proposed post-goof-trimmed LASSO in the mean regression problem. Our analysis builds upon the ideas in [1], who established the properties of post-penalized procedures for the related, but different, problem of median regression. Our analysis also builds on the fundamental results of [2] and the other works cited above that established the properties of the first-step LASSO-type estimators. An important ingredient in our analysis is a new sparsity bound on the dimension of the model selected by LASSO, which guarantees that this dimension is at most of the same order as the dimension of the "true" model. This result builds on some inequalities for sparse eigenvalues and reasoning previously given in [1] in the context of median regression. Our sparsity bounds for LASSO improve upon the analogous bounds in [2] and are comparable to the bounds in [20] obtained under a larger penalty level. We also rely on maximal inequalities in [20] to provide primitive conditions for the sharp sparsity bounds to hold.

We organize the remainder of the paper as follows. In Section 2, we review some benchmark results of [2] for LASSO, albeit with a slightly improved choice of penalty, and model selection results of [11, 13, 21]. In Section 3, we present a generic rate result on post-penalized estimators. In Section 4, we present a generic rate result for post-trimmed-estimators, where trimming can

be traditional or based on goodness-of-fit. In Section 5, we apply our generic results to the post-LASSO and the post-trimmed LASSO estimators. In Section 6 we present the results of our computational experiments.

Notation. In what follows, all parameter values are indexed by the sample size n, but we omit the index whenever this does not cause confusion. We use the notation $(a)_+ = \max\{a, 0\}$, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. The ℓ_2-norm is denoted by $\|\cdot\|$ and the ℓ_0-norm $\|\cdot\|_0$ denotes the number of non-zero components of a vector. Given a vector $\delta \in \mathbb{R}^p$, and a set of indices $T \subset \{1, \ldots, p\}$, we denote by δ_T the vector in which $\delta_{Tj} = \delta_j$ if $j \in T$, $\delta_{Tj} = 0$ if $j \notin T$. We also use standard notation in the empirical process literature, $\mathbb{E}_n[f] = \mathbb{E}_n[f(z_i)] = \sum_{i=1}^n f(z_i)/n$, and $\mathbb{G}_n(f) = \sum_{i=1}^n (f(z_i) - E[f(z_i)])/\sqrt{n}$. We use the notation $a \lesssim b$ to denote $a \leq cb$ for some constant $c > 0$ that does not depend on n; and $a \lesssim_P b$ to denote $a = O_P(b)$. For an event E, we say that E wp $\to 1$ when E occurs with probability approaching one as n grows.

2. LASSO as a Benchmark in Parametric and Nonparametric Models

The purpose of this section is to define the models for which we state our results and also to revisit some known results for the LASSO estimator, which we will use as a benchmark and as inputs to subsequent proofs. In particular, we revisit the fundamental rate results of [2], but with a slightly improved, data-driven penalty level.

2.1. Model 1: Parametric Model.

Let us consider the following parametric linear regression model:

$$y_i = x_i'\beta_0 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad \beta_0 \in \mathbb{R}^p, \quad i = 1, \ldots, n$$
$$T = \text{support}(\beta_0) \text{ has } s \text{ elements where } s < n, \text{but } p > n,$$

where T is unknown, and regressors $X = [x_1, \ldots, x_n]'$ are fixed and normalized so that $\widehat{\sigma}_j^2 = \mathbb{E}_n[x_{ij}^2] = 1$ for all $j = 1, \ldots, p$.

Given the large number of regressors $p > n$, some regularization is required in order to avoid overfitting the data. The LASSO estimator [16] is one way to achieve this regularization. Specifically, define

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \widehat{Q}(\beta) + \frac{\lambda}{n}\|\beta\|_1, \quad \text{where} \quad \widehat{Q}(\beta) = \mathbb{E}_n[y_i - x_i'\beta]^2. \tag{2.1}$$

Our goal is to revisit convergence results for $\widehat{\beta}$ in the *prediction* (pseudo) norm,

$$\|\delta\|_{2,n} = \sqrt{\mathbb{E}_n[x_i'\delta]^2}.$$

The key quantity in the analysis is the gradient at the true value:

$$S = 2\mathbb{E}_n[x_i \epsilon_i].$$

This gradient is the effective "noise" in the problem. Indeed, for $\delta = \widehat{\beta} - \beta_0$, we have by the Hölder inequality

$$\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) - \|\delta\|_{2,n}^2 = 2\mathbb{E}_n[\epsilon_i x_i'\delta] \geq -\|S\|_\infty \|\delta\|_1. \tag{2.2}$$

Thus, $\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0)$ provides noisy information about $\|\delta\|_{2,n}^2$, and the amount of noise is controlled by $\|S\|_\infty \|\delta\|_1$. This noise should be dominated by the penalty, so that the rate of convergence can be deduced from a relationship between the penalty term and the quadratic term $\|\delta\|_{2,n}^2$.

This reasoning suggests choosing λ so that

$$\lambda \geq cn\|S\|_\infty, \quad \text{for some fixed} \quad c > 1.$$

However this choice is not feasible, since we do not know S. We propose setting

$$\lambda = c \cdot \Lambda(1 - \alpha | X) \tag{2.3}$$

where $\Lambda(1 - \alpha|X)$ is the $(1-\alpha)$-quantile of $n\|S\|_\infty$, so that for this choice

$$\lambda \geq cn\|S\|_\infty \text{ with probability at least } 1 - \alpha. \tag{2.4}$$

Note that the quantity $\Lambda(1 - \alpha|X)$ is easily computed by simulation. We refer to this choice of λ as the data-driven choice, reflecting the dependence of the choice on the design matrix X.

Comment 2.1 (Data-driven choice vs standard choice.). The standard choice of λ employs

$$\lambda = c \cdot \sigma A \sqrt{2n \log p}, \tag{2.5}$$

where $A \geq 1$ is a constant that does not depend on X, chosen so that (2.4) holds no matter what X is. Note that $\sqrt{n}\|S\|_\infty$ is a maximum of $N(0, \sigma^2)$ variables, which are correlated if columns of X are correlated, as they typically are in the sample. In order to compute A, the standard choice uses the conservative assumption that these variables are uncorrelated. When the variables are highly correlated, the standard choice (2.5) becomes quite conservative and may be too large. The X-dependent choice of penalty (2.3) takes advantage of the in-sample correlations induced by the design matrix and yields smaller penalties. To illustrate this point, we simulated many designs X by drawing $\tilde{x}_i$ as i.i.d. from $N(0, \Sigma)$, and defining $x_{ij} = \tilde{x}_{ij}/\sqrt{\mathbb{E}_n[\tilde{x}_{ij}^2]}$, with $\Sigma_{jj} = 1$, and varying correlations Σ_{jk} for $j \neq k$ among three design options: 0, $\rho^{|j-k|}$, or ρ. We then

computed X-dependent penalty levels (2.3). Figure 1 plots the sorted realized values of the X-dependent λ and illustrates the impact of in-sample correlation on these values. As expected, for a fixed confidence level $1-\alpha$, the more correlated the regressors are, the smaller the data-driven penalty (2.3) is relative to the standard conservative choice (2.5).

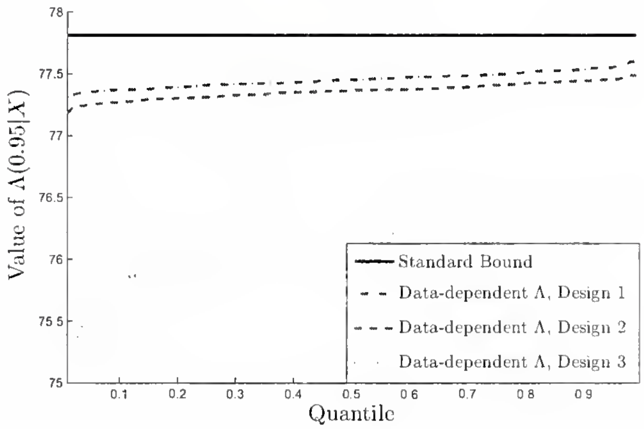

FIGURE 1. Realized values of $\Lambda(0.95|X)$ sorted in increasing order. X is drawn by generating x_i as i.i.d. $N(0,\Sigma)$, where for $j \neq k$ design 1 has $\Sigma_{jk}=0$, design 2 has $\Sigma_{jk}=(1/2)^{|j-k|}$, and design 3 has $\Sigma_{jk}=1/2$. We used $n=100$, $p=500$ and $\sigma^2=1$. For each design 100 design matrices were drawn.

Under (2.3), $\delta=\widehat{\beta}-\beta_0$ will obey the following "restricted condition" with probability at least $1-\alpha$:

$$\|\delta_{T^c}\|_1 \leq \bar{c}\|\delta_T\|_1, \quad \text{where } \bar{c} := \frac{c+1}{c-1}. \tag{2.6}$$

Therefore, in order to get convergence rates in the prediction norm $\|\delta\|_{2,n}=\sqrt{\mathbb{E}_n[x_i'\delta]^2}$, we consider the following modulus of continuity between the penalty and the prediction norm:

$$\text{RE.1}(c) \quad \kappa_1(T) := \min_{\|\delta_{T^c}\|_1 \leq \bar{c}\|\delta_T\|_1, \delta \neq 0} \frac{\sqrt{s}\|\delta\|_{2,n}}{\|\delta_T\|_1},$$

where $\kappa_1(T)$ can depend on n. In turn, the convergence rate in the usual Euclidian norm $\|\delta\|$ is determined by the following modulus of continuity between the prediction norm and the Euclidian norm:

$$\text{RE.2 }(c) \quad \kappa_2(T) := \min_{\|\delta_{T^c}\|_1 \leq \bar{c}\|\delta_T\|_1, \delta \neq 0} \frac{\|\delta\|_{2,n}}{\|\delta\|},$$

where $\kappa_2(T)$ can depend on n. Conditions RE.1 and RE.2 are simply variants of the original restricted eigenvalue conditions imposed in Bickel, Ritov and Tsybakov [2]. In what follows, we suppress dependence on T whenever convenient.

Lemma 1 below states the rate of convergence in the prediction norm under a data-driven choice of penalty.

Lemma 1 (Essentially in Bickel, Ritov, and Tsybakov [2]). *If $\lambda \geq cn\|S\|_\infty$, then*

$$\|\widehat{\beta} - \beta_0\|_{2,n} \leq \left(1 + \frac{1}{c}\right) \frac{\lambda\sqrt{s}}{n\kappa_1}.$$

Under the data-driven choice (2.3), we have with probability at least $1 - \alpha$

$$\|\widehat{\beta} - \beta_0\|_{2,n} \leq (1 + c)\, \frac{\sqrt{s}}{n\kappa_1} \Lambda(1 - \alpha|X),$$

where $\Lambda(1 - \alpha|X) \leq \sigma\sqrt{2n\log(p/\alpha)}$.

Thus, provided κ_1 is bounded away from zero, LASSO estimates the regression function at nearly the rate $\sqrt{s/n}$ (achievable when the true model T is known) with probability at least $1 - \alpha$. Since $\delta = \widehat{\beta} - \beta_0$ obeys the restricted condition with probability at least $1 - \alpha$, the rate in the Euclidian norm immediately follows from the relation

$$\|\widehat{\beta} - \beta_0\|_2 \leq \|\widehat{\beta} - \beta_0\|_{2,n}/\kappa_2, \tag{2.7}$$

which also holds with probability at least $1 - \alpha$. Thus, if κ_2 is also bounded away from zero, LASSO estimates the regression coefficients at a near $\sqrt{s/n}$ rate with probability at least $1 - \alpha$. Note that the $\sqrt{s/n}$ rate is not the oracle rate in general, but under some further conditions stated in Section 2.3, namely when the parametric model is the oracle model, this rate is an oracle rate.

2.2. **Model 2: Nonparametric model.** Next we consider the nonparametric model given by

$$y_i = f(z_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, ...n,$$

where y_i are the outcomes, z_i are vectors of fixed regressors, and ϵ_i are disturbances. For $x_i = p(z_i)$, where $p(z_i)$ is a p-vector of transformations of z_i and any conformable vector β_0, and $f_i = f(z_i)$, we can rewrite

$$y_i = x_i'\beta_0 + u_i, \quad u_i = r_i + \epsilon_i, \quad \text{where } r_i := f_i - x_i'\beta_0.$$

Next we choose our target or "true" β_0, with the corresponding cardinality of its support $s = \|\beta_0\|_0 = |T|$ as any solution to the following "oracle" risk minimization problem:

$$\min_{0 \leq k \leq p \wedge n} \min_{\|\beta\|_0 \leq k} \mathbb{E}_n[(x_i'\beta - f_i)^2] + \sigma^2 \frac{k}{n}. \tag{2.8}$$

Letting

$$c_s^2 := \mathbb{E}_n[r_i^2] = \mathbb{E}_n[(x_i'\beta_0 - f_i)^2]$$

denote the error from approximating f_i by $x_i'\beta_0$, then $c_s^2 + \sigma^2 s/n$ is the optimal value of (2.8). In order to simplify exposition, we focus some results and discussions on the case where the following holds:

$$c_s^2 \leq K\sigma^2 s/n \tag{2.9}$$

with $K = 1$ which covers most cases of interest. Alternatively, we could consider an arbitrary K which does not affect the results' modulo constants.

Note that $c_s^2 + \sigma^2 s/n$ is the the expected estimation error $E[\mathbb{E}_n[f_i - x_i'\widehat{\beta}^o]^2]$ of the (infeasible) oracle estimator $\widehat{\beta}^o$ that minimizes the expected estimation error among all k-sparse least square estimators, by searching for the best k-dimensional model and then choosing k to balance approximation error with the sampling error, which the oracle knows how to compute. The rate of convergence of the oracle estimator $\sqrt{c_s^2 + \sigma^2 s/n}$ becomes an ideal goal for the rate of convergence, and in general can be achieved only up to logarithmic terms in most cases (see Donoho and Jonstone [7] and Rigollet and Tsybakov [14]), except under very special circumstances, such as when it becomes possible to perform perfect model selection. Finally, note that when the approximation error, c_s, is zero the oracle model becomes the parametric model of the previous section where we had $r_t = 0$.

Next we state a rate of convergence in the prediction norm under the data-driven choice of penalty.

Lemma 2 (Essentially in Bickel, Ritov, and Tsybakov [2]). *If $\lambda \geq cn\|S\|_\infty$, then*

$$\|\widehat{\beta} - \beta_0\|_{2,n} \leq \left(1 + \frac{1}{c}\right) \frac{\lambda\sqrt{s}}{n\kappa_1} + 2c_s.$$

Under the data-driven choice (2.3), we have with probability at least $1 - \alpha$

$$\|\widehat{\beta} - \beta_0\|_{2,n} \leq (1 + c) \frac{\sqrt{s}}{n\kappa_1} \Lambda(1 - \alpha|X) + 2c_s,$$

where $\Lambda(1 - \alpha|X) \leq \sigma\sqrt{2n\log(p/\alpha)}$.

Thus, provided κ_1 is bounded away from zero, LASSO estimates the regression function at a near-oracle rate with probability at least $1 - \alpha$. Furthermore, the bound on empirical risk follows from the triangle inequality:

$$\sqrt{\mathbb{E}_n[x_i'\widehat{\beta} - f_i]^2} \leq \|\widehat{\beta} - \beta_0\|_{2,n} + c_s. \tag{2.10}$$

2.3. **Model Selection Properties.** The primary results we develop do not require the first-step estimators like LASSO to perfectly select the true model. In fact, we are specifically interested in the most common cases where these estimators do not perfectly select the true model. For these cases, we will prove that post-model selection estimators such as post-LASSO achieve near-oracle rates like those of LASSO. However, in some special cases, where perfect model selection is possible, these estimators can achieve the exact oracle rates, and thus can be even better than LASSO. The purpose of this section is to describe these very special cases where perfect model selection is possible.

In the discussion of our results on post-penalized estimators we will refer to the following model selection results for the parametric case.

Lemma 3 (Essentially in Meinshausen and Yu [13] and Lounici [11]). *1) In the parametric model, if the coefficients are well separated from zero, that is*

$$\min_{j\in T} |\beta_{0j}| > \ell + t, \qquad \text{for } t \geq \ell := \max_{j=1,\ldots,p} |\widehat{\beta}_j - \beta_{0j}|,$$

then the true model is a subset of the selected model, $T := \text{support}(\beta_0) \subseteq \widehat{T} := \text{support}(\widehat{\beta})$. Moreover T can be perfectly selected by applying trimming of level t to $\widehat{\beta}$:

$$T = \widetilde{T}(t) := \left\{ j \in \{1, \ldots, p\} \ : \ |\widehat{\beta}_j| > t \right\}.$$

2) In particular, if $\lambda \geq cn\|S\|_\infty$, then

$$\ell \leq \left(1 + \frac{1}{c}\right) \frac{\lambda\sqrt{s}}{n\kappa_1\kappa_2}.$$

3) In particular, if $\lambda \geq cn\|S\|_\infty$, and there is a $u \geq 1$ such that the design matrix satisfies $|\mathbb{E}_n[x_{ij}x_{ik}]| \leq 1/(u(1 + 2\bar{c})s)$ for all $1 \leq j < k \leq p$, then

$$\ell \leq \frac{\lambda}{n}\left(1 + \frac{2}{\sqrt{u^2 - u}}\right).$$

Thus, we see from parts 1) and 2), which follow from [13] and Lemma 2, that perfect model selection is possible under strong assumptions on the coefficients' separation away from zero. We also see from part 3), which is due to [11], that the strong separation of coefficients can be

considerably weakened in exchange for a strong assumption on the design matrix. Finally, the following extreme result also requires strong assumptions on separation of coefficients and the design matrix.

Lemma 4 (Essentially in Zhao and Yu [21]). *In the parametric model, under more restrictive conditions on the design, separation of non-zero coefficients, and penalty parameter, specified in* [21], *with a high probability*

$$T = \text{support}(\beta_0) = \widehat{T} = \text{support}(\widehat{\beta}).$$

Comment 2.2. We only review model selection in the parametric case. There are two reasons for this: first, the results stated above have been developed for the parametric case only, and extending them to nonparametric cases is outside the main focus of this paper. Second, it is clear from the stated conditions that in the nonparametric context, in order to select the oracle model T perfectly, the oracle models have to be either (a) parametric (i.e. $c_s = 0$) or (b) very close to parametric (with c_s much smaller than $\sigma^2 s/n$) and satisfy other strong conditions similar to those stated above. Since we only argue that post-LASSO and related estimators are as good as LASSO and can be strictly better only in some cases, it suffices to demonstrate the latter for case (a). Moreover, if oracle performance is achieved for case (a), then by continuity of empirical risk with respect to the underlying model, the oracle performance should extend to a neighborhood of case (a), which is case (b).

3. A Generic Result on Post-Model Selection Estimators

Let $\widehat{\beta}$ be any first-step estimator acting as a model selection device and denote by

$$\widehat{T} := \text{support}(\widehat{\beta})$$

the model selected by this estimator; we assume $|\widehat{T}| \leq n$ throughout. Define the post-model selection estimator as

$$\widetilde{\beta} = \arg \min_{\beta_{\widehat{T}^c}=0} \widehat{Q}(\beta). \tag{3.11}$$

If model selection works perfectly (as it will under some rather stringent conditions), then this estimator is simply the oracle estimator and its properties are well known. However, of more interest is the case when model selection does not work perfectly, as occurs for many designs of interest in applications. In this section we derive a generic result on the performance of any post-model selection estimator.

In order to derive rates, we need the following minimal restricted sparse eigenvalue

$$\text{RSE.1}(m) \quad \widetilde{\kappa}(m)^2 := \min_{\|\delta_{T^c}\|_0 \leq m, \delta \neq 0} \frac{\|\delta\|_{2,n}^2}{\|\delta\|^2}$$

as well as the following maximal restricted sparse eigenvalue

$$\text{RSE.2}(m) \quad \phi(m) := \max_{\|\delta_{T^c}\|_0 \leq m, \delta \neq 0} \frac{\|\delta\|_{2,n}^2}{\|\delta\|^2}$$

where m is the restriction on the number of non-zero components outside the support T. It will be convenient to define the following condition number associated with the sample design matrix:

$$\mu_m = \frac{\sqrt{\phi(m)}}{\widetilde{\kappa}(m)}. \tag{3.12}$$

The following theorem establishes bounds on the prediction error of a generic second-step estimator.

Theorem 1 (Performance of a generic second-step estimator). *In either the parametric model or the nonparametric model, let $\widehat{\beta}$ be any first-step estimator with support $\widehat{T}$, define*

$$B_n := \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) \quad \text{and} \quad C_n := \widehat{Q}(\beta_{0\widehat{T}}) - \widehat{Q}(\beta_0),$$

and let $\widetilde{\beta}$ the second-step estimator. For any $\varepsilon > 0$, there is a constant K_ε independent of n such that with probability at least $1 - \varepsilon$, we have that for $\widehat{m} := |\widehat{T} \setminus T|$

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \leq K_\varepsilon \sigma \sqrt{\frac{\widehat{m} \log p + (\widehat{m} + s) \log \mu_{\widehat{m}}}{n}} + 2c_s + \sqrt{(B_n)_+ \wedge (C_n)_+},$$

where $c_s = 0$ in the parametric model. Furthermore, B_n and C_n obey bounds (3.13) stated below.

The following lemma bounds B_n and C_n, although in many cases we can bound B_n by other means, as we shall do in the LASSO case.

Lemma 5 (Generic control of B_n and C_n). *Let $\widehat{m} = |\widehat{T} \setminus T|$ be the number of wrong regressors selected and $\widehat{k} = |T \setminus \widehat{T}|$ be the number of correct regressors missed. For any $\varepsilon > 0$ there is a constant K_ε independent of n such that with probability at least $1 - \varepsilon$,*

$$\begin{aligned} B_n &\leq \|\widehat{\beta} - \beta_0\|_{2,n}^2 + \left[K_\varepsilon \sigma \sqrt{\frac{\widehat{m} \log p + (\widehat{m} + s) \log \mu_{\widehat{m}}}{n}} + 2c_s \right] \|\widehat{\beta} - \beta_0\|_{2,n} \\ C_n &\leq 1\{T \not\subseteq \widehat{T}\} \left(\|\beta_{0\widehat{T}^c}\|_{2,n}^2 + \left[K_\varepsilon \sigma \sqrt{\frac{\log \binom{s}{\widehat{k}} + \widehat{k} \log \mu_0}{n}} + 2c_s \right] \|\beta_{0\widehat{T}^c}\|_{2,n} \right). \end{aligned} \tag{3.13}$$

Three implications of Theorem 1 are worth noting. Firstly and most importantly, note that the bounds on the prediction norm stated in Theorem 1 and Lemma 5 apply to any generic post-model selection estimator, provided we can bound both the rate of convergence $\|\widehat{\beta}-\beta_0\|_{2,n}$ of the first-step estimator and $\widehat{m}$, the number of wrong regressors selected by the first-step estimator.

Secondly, note that if the selected model contains the true model, $T \subseteq \widehat{T}$, then we have $(B_n)_+ \wedge (C_n)_+ = C_n = 0$, and B_n does not affect the rate at all, and the performance of the second-step estimator is dictated by the sparsity $\widehat{m}$ of the first-step estimator, which controls the magnitude of the empirical errors. Otherwise, if the selected model fails to contain the true model, that is, $T \not\subseteq \widehat{T}$, the performance of the second-step estimator is determined by both the sparsity $\widehat{m}$ and the minimum between B_n and C_n. Intuitively, B_n measures the in-sample goodness-of-fit (or loss-of-fit) induced by the first-step estimator relative to the "true" parameter value β_0, and C_n measures the in-sample loss-of-fit induced by truncating the "true" parameter β_0 outside the selected model $\widehat{T}$.

Finally, note that rates in other norms of interest immediately follow from the following relations:

$$\sqrt{\mathbb{E}_n[x_i'\widetilde{\beta} - f_i]^2} \leq \|\widetilde{\beta} - \beta_0\|_{2,n} + c_s, \quad \|\widetilde{\beta} - \beta_0\|_2 \leq \|\widetilde{\beta} - \beta_0\|_{2,n}/\widetilde{\kappa}(\widehat{m}), \tag{3.14}$$

where $\widehat{m} = |\widehat{T} \setminus T|$.

The proof of Theorem 1 and Lemma 5 relies on the sparsity-based control of the empirical error provided by the following lemma.

Lemma 6 (Sparsity-based control of noise). *1) For any $\varepsilon > 0$, there is a constant K_ε independent of n such that with probability at least $1-\varepsilon$,*

$$|\widehat{Q}(\beta_0+\delta) - \widehat{Q}(\beta_0) - \|\delta\|_{2,n}^2| \leq K_\varepsilon \sigma \sqrt{\frac{m\log p + (m+s)\log \mu_m}{n}}\|\delta\|_{2,n} + 2c_s\|\delta\|_{2,n},$$

uniformly for all $\delta \in \mathbb{R}^p$ such that $\|\delta_{T^c}\|_0 \leq m$, and uniformly over $m \leq n$,

where $c_s = 0$ in the parametric model. 2) Furthermore, with at least the same probability,

$$|\widehat{Q}(\beta_{0\widetilde{T}}) - \widehat{Q}(\beta_0) - \|\beta_{0\widetilde{T}^c}\|_{2,n}^2| \leq K_\varepsilon \sigma \sqrt{\frac{\log\binom{s}{k} + k\log \mu_0}{n}}\|\beta_{0\widetilde{T}^c}\|_{2,n} + 2c_s\|\beta_{0\widetilde{T}^c}\|_{2,n},$$

uniformly for all $\widetilde{T} \subset T$ such that $|T \setminus \widetilde{T}| = k$, and uniformly over $k \leq s$,

where $c_s = 0$ in the parametric model.

The proof of the lemma in turn relies on the following maximal inequality, which we state as a separate theorem since it may be of independent interest. The proof of the theorem involves the use of Samorodnitsky-Talagrand's inequality.

Theorem 2 (Maximal inequality for a collection of empirical processes). *Let $\epsilon_i \sim N(0,\sigma^2)$ be independent for $i = 1,\dots,n$, and for $m = 1,\dots,n$ define*

$$e_n(m,\eta) := \sigma 2\sqrt{2}\left(\sqrt{\log\binom{p}{m}} + \sqrt{(m+s)\log(D\mu_m)} + \sqrt{(m+s)\log(1/\eta)}\right)$$

for any $\eta \in (0,1)$ and some universal constant D. Then

$$\sup_{\|\delta_{T^c}\|_0 \le m, \|\delta\|_{2,n}>0} \left|\mathbb{G}_n\left(\frac{\epsilon_i x_i'\delta}{\|\delta\|_{2,n}}\right)\right| \le e_n(m,\eta), \text{ for all } m \le n,$$

with probability at least $1 - \eta e^{-s}/(1-1/e)$.

Proof. Step 0. Note that we can restrict the supremum over $\|\delta\| = 1$ since the function is homogenous of degree zero.

Step 1. For each non-negative integer $m \le n$, and each set $\widetilde{T} \subset \{1,\dots,p\}$, with $|\widetilde{T}\setminus T| \le m$, define the class of functions

$$\mathcal{G}_{\widetilde{T}} = \{\epsilon_i x_i'\delta/\|\delta\|_{2,n} : \text{support}(\delta) \subseteq \widetilde{T}, \|\delta\| = 1\}. \tag{3.15}$$

Also define

$$\mathcal{F}_m = \{\mathcal{G}_{\widetilde{T}} : \widetilde{T} \subset \{1,\dots,p\} : |\widetilde{T}\setminus T| \le m\}.$$

It follows that

$$P\left(\sup_{f\in\mathcal{F}_m} |\mathbb{G}_n(f)| \ge e_n(m,\eta)\right) \le \binom{p}{m} \max_{|\widetilde{T}\setminus T|\le m} P\left(\sup_{f\in\mathcal{G}_{\widetilde{T}}} |\mathbb{G}_n(f)| \ge e_n(m,\eta)\right). \tag{3.16}$$

We apply Samorodnitsky-Talagrand's inequality (Proposition A.2.7 in van der Vaart and Wellner [18]) to bound the right hand side of (3.16). Let

$$\rho(f,g) := \sqrt{E_\epsilon[\mathbb{G}_n(f) - \mathbb{G}_n(g)]^2} = \sqrt{E_\epsilon\mathbb{E}_n[(f-g)^2]}$$

for $f,g \in \mathcal{G}_{\widetilde{T}}$; by Step 2 below, the covering number of $\mathcal{G}_{\widetilde{T}}$ with respect to ρ obeys

$$N(\varepsilon, \mathcal{G}_{\widetilde{T}}, \rho) \le (6\sigma\mu_m/\varepsilon)^{m+s}, \text{ for each } 0 < \varepsilon \le \sigma, \tag{3.17}$$

and $\sigma^2(\mathcal{G}_{\widetilde{T}}) := \max_{f\in\mathcal{G}_{\widetilde{T}}} E[\mathbb{G}_n(f)]^2 = \sigma^2$. Then, by Samorodnitsky-Talagrand's inequality

$$P\left(\sup_{f\in\mathcal{G}_{\widetilde{T}}} |\mathbb{G}_n(f)| \ge e_n(m,\eta)\right) \le \left(\frac{D\sigma\mu_m e_n(m,\eta)}{\sqrt{m+s}\,\sigma^2}\right)^{m+s} \bar{\Phi}(e_n(m,\eta)/\sigma)$$

for some universal constant $D \geq 1$. For $e_n(m,\eta)$ defined in the statement of the theorem, it follows that

$$P\left(\sup_{f\in\mathcal{G}_{\widetilde{T}}}|\mathbb{G}_n(f)| \geq e_n(m,\eta)\right) \leq \eta e^{-m-s}/\binom{p}{m}.$$

Then,

$$\begin{aligned} P\left(\sup_{f\in\mathcal{F}_m}|\mathbb{G}_n(f)| > e_n(m,\eta), \exists m \leq n\right) &\leq \sum_{m=0}^{n} P\left(\sup_{f\in\mathcal{F}_m}|\mathbb{G}_n(f)| > e_n(m,\eta)\right) \\ &\leq \sum_{m=0}^{n} \eta e^{-m-s} \leq \eta e^{-s}/(1-1/e), \end{aligned}$$

which proves the claim of the theorem.

Step 2. This step establishes (3.17). For $t \in \mathbb{R}^p$ and $\widetilde{t} \in \mathbb{R}^p$, consider any two functions

$$\epsilon_i \frac{(x_i't)}{\|t\|_{2,n}} \text{ and } \epsilon_i \frac{(x_i'\widetilde{t})}{\|\widetilde{t}\|_{2,n}} \text{ in } \mathcal{G}_{\widetilde{T}}, \text{ for a given } \widetilde{T} \subset \{1,...,p\} : |\widetilde{T}\setminus T| \leq m.$$

We have that

$$\sqrt{E_\epsilon \mathbb{E}_n\left[\epsilon_i^2\left(\frac{(x_i't)}{\|t\|_{2,n}} - \frac{(x_i'\widetilde{t})}{\|\widetilde{t}\|_{2,n}}\right)^2\right]} \leq \sqrt{E_\epsilon \mathbb{E}_n\left[\epsilon_i^2 \frac{(x_i'(t-\widetilde{t}))^2}{\|t\|_{2,n}^2}\right]} + \sqrt{E_\epsilon \mathbb{E}_n\left[\epsilon_i^2\left(\frac{(x_i'\widetilde{t})}{\|t\|_{2,n}} - \frac{(x_i'\widetilde{t})}{\|\widetilde{t}\|_{2,n}}\right)^2\right]}.$$

By definition of $\mathcal{G}_{\widetilde{T}}$ in (3.15), support$(t) \subseteq \widetilde{T}$ and support$(\widetilde{t}) \subseteq \widetilde{T}$, so that support$(t-\widetilde{t}) \subseteq \widetilde{T}$, $|\widetilde{T}\setminus T| \leq m$, and $\|t\| = 1$ by (3.15). Hence by definitions RSE.1(m) and RSE.2(m),

$$E_\epsilon \mathbb{E}_n\left[\epsilon_i^2 \frac{(x_i'(t-\widetilde{t}))^2}{\|t\|_{2,n}^2}\right] \leq \sigma^2\phi(m)\|t-\widetilde{t}\|^2/\widetilde{\kappa}(m)^2, \text{ and}$$

$$\begin{aligned} E_\epsilon \mathbb{E}_n\left[\epsilon_i^2\left(\frac{(x_i'\widetilde{t})}{\|t\|_{2,n}} - \frac{(x_i'\widetilde{t})}{\|\widetilde{t}\|_{2,n}}\right)^2\right] &= E_\epsilon \mathbb{E}_n\left[\epsilon_i^2 \frac{(x_i'\widetilde{t})^2}{\|\widetilde{t}\|_{2,n}^2}\left(\frac{\|\widetilde{t}\|_{2,n} - \|t\|_{2,n}}{\|t\|_{2,n}}\right)^2\right] \\ &= \sigma^2\left(\frac{\|\widetilde{t}\|_{2,n} - \|t\|_{2,n}}{\|t\|_{2,n}}\right)^2 \leq \sigma^2\|\widetilde{t}-t\|_{2,n}^2/\|t\|_{2,n}^2 \\ &\leq \sigma^2\phi(m)\|\widetilde{t}-t\|^2/\widetilde{\kappa}(m)^2. \end{aligned}$$

Thus

$$\sqrt{E_\epsilon \mathbb{E}_n\left[\epsilon_i^2\left(\frac{(x_i't)}{\|t\|_{2,n}} - \frac{(x_i'\widetilde{t})}{\|\widetilde{t}\|_{2,n}}\right)^2\right]} \leq 2\sigma\|t-\widetilde{t}\|\sqrt{\phi(m)}/\widetilde{\kappa}(m) = 2\sigma\mu_m\|t-\widetilde{t}\|.$$

Then the bound (3.17) follows from the bound in [18] page 94 with $R = 2\sigma\mu_m$ for any $\varepsilon \leq \sigma$.

□

Proof of Theorem 1. Let $\widetilde{\delta} := \widetilde{\beta} - \beta_0$. By definition of the second-step estimator, it follows that $\widehat{Q}(\widetilde{\beta}) \leq \widehat{Q}(\widehat{\beta})$ and $\widehat{Q}(\widetilde{\beta}) \leq \widehat{Q}(\beta_{0\widehat{T}})$. Thus,

$$\widehat{Q}(\widetilde{\beta}) - \widehat{Q}(\beta_0) \leq \left(\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0)\right) \wedge \left(\widehat{Q}(\beta_{0\widehat{T}}) - \widehat{Q}(\beta_0)\right) \leq B_n \wedge C_n.$$

By Lemma 6 part (1), for any $\varepsilon > 0$ there exists a constant K_ε such that with probability at least $1 - \varepsilon$:

$$|\widehat{Q}(\widetilde{\beta}) - \widehat{Q}(\beta_0) - \|\widetilde{\delta}\|_{2,n}^2| \leq A_{\varepsilon,n}\|\widetilde{\delta}\|_{2,n} + 2c_s\|\widetilde{\delta}\|_{2,n}$$

where

$$A_{\varepsilon,n} := K_\varepsilon \sigma \sqrt{(\widehat{m} \log p + (\widehat{m} + s) \log \mu_{\widehat{m}})/n}.$$

Combining these relations we obtain the inequality

$$\|\widetilde{\delta}\|_{2,n}^2 - A_{\varepsilon,n}\|\widetilde{\delta}\|_{2,n} - 2c_s\|\widetilde{\delta}\|_{2,n} \leq B_n \wedge C_n,$$

solving which we obtain the stated result:

$$\|\widetilde{\delta}\|_{2,n} \leq A_{\varepsilon,n} + 2c_s + \sqrt{(B_n)_+ \wedge (C_n)_+}.$$

□

4. A Generic Result on Post-Trimmed Estimators

In this section we investigate post-trimmed estimators that arise from applying unpenalized least squares in the second-step to the models selected by trimmed estimators in the first step. Formally, given a first-step estimator $\widehat{\beta}$, we define its trimmed support at level $t \geq 0$ as

$$\widetilde{T}(t) := \{j \in \{1, \ldots, p\} : |\widehat{\beta}_j| > t\}.$$

We then define the post-trimmed estimator as

$$\widetilde{\beta}^t = \arg \min_{\beta_{\widetilde{T}^c(t)}=0} \widehat{Q}(\beta). \tag{4.18}$$

The traditional trimming scheme sets the trimming threshold $t \geq \ell = \max_{1 \leq j \leq p} |\widehat{\beta}_j - \beta_{0j}|$, so that to trim all small coefficient estimates smaller than the uniform estimation error ℓ. As discussed in Section 2.3, this method is particularly appealing in parametric models in which the non-zero components are well separated from zero, where it acts as a very effective model selection device. Unfortunately, this scheme may perform poorly in parametric models with true coefficients not well separated from zero and in nonparametric models. Indeed, even in

parametric models with many small but non-zero true coefficients, trimming the estimates too aggressively may result in large goodness-of-fit losses, and consequently in slow rates of convergence and even inconsistency for the second-step estimators. This issue directly motivates our new goodness-of-fit based trimming method, which trims small coefficient estimates as much as possible subject to maintaining a certain goodness-of-fit level. Unlike traditional trimming, our new method is completely data-driven, which makes it appealing for practice. Moreover, our method is at least as good as LASSO or post-LASSO theoretically, but performs better than both of these methods in a wide range of experiments, practically. In the remainder of the section we present generic performance bounds for both the new method and the traditional trimming method.

4.1. **Goodness-of-Fit Trimming.** Here we propose a trimming method that selects the trimming level t based on the goodness-of-fit of the post-trimmed estimator. Let $\gamma \leq 0$ denoted the maximal allowed loss (gain) in goodness-of-fit (goof) relative to the first-step estimator. We define the goof-trimming threshold t_γ as the solution to

$$t_\gamma := \max_{t \geq 0}\{t : \widehat{Q}(\widetilde{\beta}^t) - \widehat{Q}(\widehat{\beta}) \leq \gamma\}. \tag{4.19}$$

Then we define the selected model and the post-goof-trimmed estimators as:

$$\widetilde{T} := \widetilde{T}(t_\gamma) \quad \text{and} \quad \widetilde{\beta} := \widetilde{\beta}^{t_\gamma}. \tag{4.20}$$

Our construction (4.19) and (4.20) selects the most aggressive trimming threshold subject to maintaining a certain level of goodness-of-fit as measured by the least squares criterion function. Note that we can compute the data-driven trimming threshold (4.19) very effectively using a binary search procedure described below.

Theorem 3 (Performance of a generic post-goof-trimmed estimator). *In either the parametric or the nonparametric model, let $\widehat{\beta}$ be any first-step estimator, $\widetilde{m} := |\widetilde{T} \setminus T|$, and $B_n := \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0)$ and $C_n := \widehat{Q}(\beta_{0\widetilde{T}}) - \widehat{Q}(\beta_0)$. For any $\varepsilon > 0$, there is a constant K_ε independent of n such that with probability at least $1 - \varepsilon$*

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \leq K_\varepsilon \sigma \sqrt{\frac{\widetilde{m}\log p + (\widetilde{m} + s)\log \mu_{\widetilde{m}}}{n}} + 2c_s + \sqrt{(\gamma + B_n)_+ \wedge (C_n)_+}, \tag{4.21}$$

where $c_s = 0$ in the parametric model. Furthermore, B_n and C_n obey bounds (3.13) stated earlier, with $\widehat{T} = \widetilde{T}$.

Note that the bounds on the prediction norm stated in Theorem 3 and equation of (3.13) in Lemma 5 apply to any generic post-goof-trimmed estimator, provided we can bound both

the rate of convergence $\|\widehat{\beta} - \beta_0\|_{2,n}$ of the first-step estimator and $\widetilde{m}$, the number of wrong regressors selected by the trimmed first-step estimator. For the purpose of obtaining rates, we can often use the bound $\widetilde{m} \leq \widehat{m}$, where $\widehat{m}$ is the number of wrong regressors selected by the first-step estimator, provided the bounds on $\widehat{m}$ are tight, as, for example, in the case of LASSO. Of course, $\widetilde{m}$ is potentially much smaller than $\widehat{m}$, resulting in a smaller variance for the post-goof-trimmed estimator. For instance, in the case of LASSO, we can even have $\widetilde{m} = 0$, if the conditions of Lemma 3 on perfect model selection in the parametric model hold with the threshold $t = t_\gamma$.

Also, note that if the selected model contains the true model, that is $T \subseteq \widetilde{T}$, then we have $(B_n)_+ \wedge (C_n)_+ = C_n = 0$, and these terms drop out of the rate. Lemma 3 provides sufficient conditions for this to hold for the given threshold $t = t_\gamma$. Otherwise, if the selected model fails to contain the true model, that is, $T \not\subseteq \widetilde{T}$, the performance of the second-step estimator is determined by both $\widetilde{m}$ and $B_n \wedge C_n$.

Comment 4.1 (Recommended choice of γ). A nice feature of the theorem above is that it allows for a wide range of choices of γ. The simplest choice with good theoretical guarantees is

$$\gamma = 0,$$

which requires there to be no loss of fit relative to the first-step estimator. We can also use any (feasible) $\gamma \leq 0$, since a negative γ actually requires the second-step estimator to gain fit relative to the first-step estimator. This makes sense, since the first-step estimator can suffer from a large regularization bias. Consequently, our recommended data-driven choice is

$$\gamma = \frac{\widehat{Q}(\widetilde{\beta}^0) - \widehat{Q}(\widehat{\beta})}{2} < 0, \tag{4.22}$$

where $\widetilde{\beta}^0$ is the post-trimmed estimator for $t = 0$. The theoretical guarantees of this choice are comparable to that of $\gamma = 0$, but this proposal led to the best performance in our computational experiments. Note that if we could set $\gamma + B_n = 0$, which is not practical and not always feasible, we would eliminate the second term in the rate bound (4.21). Since $B_n \approx \widehat{Q}(\widehat{\beta}) - \sigma^2 > 0$, if $\widehat{\beta}$ has a substantial regularization bias, then we have $\gamma < 0$. Although this choice is not available in general, it provides a simple rationale for choosing $\gamma < 0$ as we did in (4.22).

Comment 4.2 (Efficient computation of t_γ). For any γ, we can compute the value t_γ by a binary search over t. Since there are at most $|\widehat{T}|$ possible relevant values of t, we can compute t_γ exactly by running at most $\lceil \log_2 |\widehat{T}| \rceil$ unpenalized least squares problems.

Proof of Theorem 3. Let $\widetilde{\delta} := \widetilde{\beta} - \beta_0$. By definition $\widehat{Q}(\widetilde{\beta}) \leq \widehat{Q}(\widehat{\beta}) + \gamma$, so that

$$\widehat{Q}(\widetilde{\beta}) - \widehat{Q}(\beta_0) \leq \gamma + \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) = \gamma + B_n.$$

On the other hand, since $\widetilde{\beta}$ is a minimizer of $\widehat{Q}$ over the support $\widetilde{T}$, $\widehat{Q}(\widetilde{\beta}) \leq \widehat{Q}(\beta_{0\widetilde{T}})$ so that

$$\widehat{Q}(\widetilde{\beta}) - \widehat{Q}(\beta_0) \leq \widehat{Q}(\beta_{0\widetilde{T}}) - \widehat{Q}(\beta_0) = C_n.$$

By Lemma 6 part (1), for any $\varepsilon > 0$, there is a constant K_ε such that with probability at least $1 - \varepsilon$

$$\|\widetilde{\delta}\|_{2,n}^2 - A_{\varepsilon,n}\|\widetilde{\delta}\|_{2,n} - 2c_s\|\widetilde{\delta}\|_{2,n} \leq \widehat{Q}(\widetilde{\beta}) - \widehat{Q}(\beta_0),$$

where

$$A_{\varepsilon,n} := K_\varepsilon \sigma \sqrt{(\widetilde{m} \log p + (\widetilde{m} + s) \log \mu_{\widetilde{m}})/n}.$$

Combining the inequalities gives

$$\|\widetilde{\delta}\|_{2,n}^2 - A_{\varepsilon,n}\|\widetilde{\delta}\|_{2,n} - 2c_s\|\widetilde{\delta}\|_{2,n} \leq (\gamma + B_n) \wedge C_n.$$

Solving this inequality for $\|\widetilde{\delta}\|_{2,n}$ gives the stated result. □

4.2. **Traditional Trimming.** Next we consider the traditional trimming scheme, which is based on the magnitude of the estimated coefficients. Given the first-step estimator $\widehat{\beta}$, define the trimmed first-step estimator $\widehat{\beta}_t$ by setting $\widehat{\beta}_{tj} = \widehat{\beta}_j 1\{|\widehat{\beta}_j| \geq t\}$ for $j = 1, ..., p$. Finally define the selected model and the post-trimmed estimator as

$$\widetilde{T} = \widetilde{T}(t) \quad \text{and} \quad \widetilde{\beta} = \widetilde{\beta}^t. \tag{4.23}$$

Let $\widetilde{m}; = |\widetilde{T} \setminus T|$ denote the components selected outside the support T, $m_t := |\widehat{T} \setminus \widetilde{T}|$ the number of trimmed components of the first-step estimator, and $\gamma_t := \|\widehat{\beta}_t - \widehat{\beta}\|_{2,n}$ the prediction norm distance from the first-step estimator $\widehat{\beta}$ to the trimmed estimator $\widehat{\beta}_t$.

Theorem 4 (Performance of a generic post-traditional-trimmed estimator). *In either the parametric or the nonparametric model, let $\widehat{\beta}$ be any first-step estimator, and let $B_n := \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0)$ and $C_n := \widehat{Q}(\beta_{0\widetilde{T}}) - \widehat{Q}(\beta_0)$. For any $\varepsilon > 0$, there is a constant K_ε independent of n such that with probability at least $1 - \varepsilon$*

$$\begin{aligned} \|\widetilde{\beta} - \beta_0\|_{2,n} &\leq K_\varepsilon \sigma \sqrt{(\widetilde{m} \log p + (\widetilde{m} + s) \log \mu_{\widetilde{m}})/n} + 2c_s + \\ &+ \sqrt{\gamma_t (K_\varepsilon \sigma G_t + 2c_s + \gamma_t + 2\|\widehat{\beta} - \beta_0\|_{2,n}) + (B_n)_+ \wedge \sqrt{(C_n)_+}}, \end{aligned}$$

where $G_t = \sqrt{m_t \log(p\mu_{m_t})/n}$, $\gamma_t \leq t\sqrt{\phi(m_t) m_t}$, and $c_s = 0$ in the parametric model. Furthermore, B_n and C_n obey bounds (3.13) stated earlier, with $\widehat{T} = \widetilde{T}$.

Note that the bounds on the prediction norm stated in Theorem 4 and equation (3.13) in Lemma 5 apply to any generic post-traditional-trimmed estimator. All components of the bounds are easily controlled, just as in the case of Theorem 3. A major determinant of the performance is γ_t which measures loss-of-fit due to trimming. If the trimming threshold is too aggressive, for example, as suggested in the model selection Lemma 3 (2), then γ_t can be very large. Indeed, in the parametric models with true coefficients not well separated from zero and in the nonparametric models, aggressive trimming can result in large goodness-of-fit losses γ_t, and consequently in very slow rates of convergence and even inconsistency for the second-step estimators. We further discuss this issue in the next section in the context of LASSO. There are of course exceptional cases where good model selection is possible. One example is the parametric model with well-separated coefficients, where $T \subseteq \widetilde{T}$ wp $\to 1$ so that $C_n = 0$ wp $\to 1$, which eliminates dependence of performance bounds on γ_t completely.

Comment 4.3 (Traditional trimming based on goodness-of-fit). We can fix some drawbacks of traditional trimming by selecting the threshold t to imply at most a specific loss of fit γ_t. For a given $\gamma_t \geq 0$, we can set $t = \max\{t : \|\widehat{\beta}_t - \widehat{\beta}\|_{2,n} \leq \gamma_t\}$. This choice uses maximal trimming subject to maintaining a certain goodness-of-fit level, as measured by the prediction norm. Our theorem above formally covers this choice. However, it is not easy to specify practical, data-driven γ_t. Our main proposal described in the previous subsection resolves just such difficulties.

Proof of Theorem 4. Let $\widetilde{\delta} := \widetilde{\beta} - \beta_0$, $\widehat{\delta}^t := \widehat{\beta}_t - \beta_0$, and $\widehat{\delta} := \widehat{\beta} - \beta_0$. By definition of the estimator, $\widehat{Q}(\widetilde{\beta}) \leq \widehat{Q}(\widehat{\beta}_t) \wedge \widehat{Q}(\beta_{0\widetilde{T}})$, so that

$$\widehat{Q}(\widetilde{\beta}) - \widehat{Q}(\beta_0) \leq \left(\widehat{Q}(\widehat{\beta}_t) - \widehat{Q}(\beta_0)\right) \wedge \left(\widehat{Q}(\beta_{0\widetilde{T}}) - \widehat{Q}(\beta_0)\right) \leq \left(\widehat{Q}(\widehat{\beta}_t) - \widehat{Q}(\widehat{\beta}) + B_n\right) \wedge C_n$$

since $B_n = \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0)$.

By Lemma 6 (1), for any $\varepsilon > 0$ there is a constant $K_{\varepsilon,1}$ such that with probability at least $1 - \varepsilon/2$

$$\|\widetilde{\delta}\|_{2,n}^2 - A_{\varepsilon,n}\|\widetilde{\delta}\|_{2,n} - 2c_s\|\widetilde{\delta}\|_{2,n} \leq \widehat{Q}(\widetilde{\beta}) - \widehat{Q}(\beta_0),$$

where

$$A_{\varepsilon,n} := K_{\varepsilon,1}\sigma\sqrt{(\widetilde{m}\log p + (\widetilde{m} + s)\log \mu_{\widetilde{m}})/n}.$$

On the other hand, we have

$$\begin{aligned} \widehat{Q}(\widehat{\beta}_t) - \widehat{Q}(\widehat{\beta}) &= \widehat{Q}(\widehat{\beta}_t) - \widehat{Q}(\beta_0) + \widehat{Q}(\beta_0) - \widehat{Q}(\widehat{\beta}) \\ &= 2\mathbb{E}_n[\epsilon_i x_i'(\widehat{\beta}_t - \widehat{\beta})] + 2\mathbb{E}_n[r_i x_i'(\widehat{\beta}_t - \widehat{\beta})] + \|\widehat{\delta}^t\|_{2,n}^2 - \|\widehat{\delta}\|_{2,n}^2. \end{aligned}$$

To bound the terms above, note first that by Theorem 2, there is a constant $K_{\varepsilon,2}$ such that with probability at least $1-\varepsilon/2$

$$|2\mathbb{E}_n[\epsilon_i x_i'(\widehat{\beta}_t - \widehat{\beta})]| \leq \sigma K_{\varepsilon,2} G_t \|\widehat{\beta}_t - \widehat{\beta}\|_{2,n},$$

and, second, by Cauchy-Schwartz $|2\mathbb{E}_n[r_i x_i'(\widehat{\beta}_t - \widehat{\beta})]| \leq 2c_s \|\widehat{\beta}_t - \widehat{\beta}\|_{2,n}$. Moreover,

$$\begin{aligned} \|\widehat{\delta}^t\|_{2,n}^2 - \|\widehat{\delta}\|_{2,n}^2 &= (\|\widehat{\delta}^t\|_{2,n} - \|\widehat{\delta}\|_{2,n})(\|\widehat{\delta}^t\|_{2,n} + \|\widehat{\delta}\|_{2,n}) \\ &\leq \|\widehat{\beta}^t - \widehat{\beta}\|_{2,n}(\|\widehat{\beta}_t - \widehat{\beta}\|_{2,n} + 2\|\widehat{\delta}\|_{2,n}). \end{aligned}$$

Combining these inequalities and using that $\gamma_t = \|\widehat{\beta}_t - \widehat{\beta}\|_{2,n}$, we obtain with probability at least $1-\varepsilon$

$$\|\widetilde{\delta}\|_{2,n}^2 - A_{\varepsilon,n}\|\widetilde{\delta}\|_{2,n} - 2c_s\|\widetilde{\delta}\|_{2,n} \leq \Big(\sigma K_{\varepsilon,2} G_t \gamma_t + 2c_s\gamma_t + \gamma_t(\gamma_t + 2\|\widehat{\delta}\|_{2,n}) + B_n\Big) \wedge C_n.$$

Thus, solving the resulting quadratic inequality for $\|\widetilde{\delta}\|_{2,n}$, we obtain

$$\|\widetilde{\delta}\|_{2,n} \leq A_{\varepsilon,n} + 2c_s + \sqrt{\Big(\sigma K_{\varepsilon,2} G_t \gamma_t + 2c_s\gamma_t + \gamma_t(\gamma_t + 2\|\widehat{\delta}\|_{2,n}) + (B_n)_+\Big) \wedge (C_n)_+},$$

which gives the stated result by taking $K_\varepsilon = K_{\varepsilon,1} \vee K_{\varepsilon,2}$. Also, note that $\gamma_t \leq t\sqrt{\phi(m_t)m_t}$ follows by the Cauchy-Schwartz inequality and the definition of $\phi(m_t)$. □

5. Post Model Selection Results for LASSO

In this section we specialize our results on post-penalized estimators to the case of LASSO being the first-step estimator. The previous generic results allow us to use sparsity bounds and rate of convergence of LASSO to derive the rate of convergence of post-penalized estimators in the parametric and nonparametric models. We also derive new sharp sparsity bounds for LASSO, which may be of independent interest.

5.1. **A new, oracle sparsity bound for LASSO.**
We begin with a preliminary sparsity bound for LASSO.

Lemma 7 (Empirical pre-sparsity for LASSO). *In either the parametric model or the nonparametric model, let $\widehat{m} = |\widehat{T} \setminus T|$ and $\lambda \geq c \cdot n\|S\|_\infty$. We have*

$$\sqrt{\widehat{m}} \leq \sqrt{s}\sqrt{\phi(\widehat{m})}\ 2\bar{c}/\kappa_1 + 3(\bar{c}+1)\sqrt{\phi(\widehat{m})}\ nc_s/\lambda,$$

where $c_s = 0$ in the parametric model.

The lemma above states that LASSO achieves the oracle sparsity up to a factor of $\phi(\widehat{m})$. The lemma above immediately yields the simple upper bound on the sparsity of the form

$$\widehat{m} \lesssim_P s\phi(n), \tag{5.24}$$

as obtained for example in [2] and [13]. Unfortunately, this bound is sharp only when $\phi(n)$ is bounded. When $\phi(n)$ diverges, for example when $\phi(n) \gtrsim_P \sqrt{\log p}$ in the Gaussian design with $p \geq 2n$, the bound is not sharp. However, for this case we can construct a sharp sparsity bound by combining the preceding pre-sparsity result with the following sub-linearity property of the restricted sparse eigenvalues.

Lemma 8 (Sub-linearity of restricted sparse eigenvalues). *For any integer $k \geq 0$ and constant $\ell \geq 1$ we have $\phi(\lceil \ell k \rceil) \leq \lceil \ell \rceil \phi(k)$.*

A version of this lemma for unrestricted eigenvalues has been previously proven in [1]. The combination of the preceding two lemmas gives the following sparsity theorem. Recall that we assume $c_s \leq \sigma\sqrt{s/n}$ and for $\alpha \leq 1/4$ we have $\Lambda(1-\alpha|X) \geq \sigma\sqrt{n}$.

Theorem 5 (Sparsity bound for LASSO under data-driven penalty). *In either the parametric model or the nonparametric model, consider the LASSO estimator with $\lambda \geq c\Lambda(1-\alpha|X)$, $\alpha \leq 1/4$, $c_s \leq \sigma\sqrt{s/n}$, and let $\widehat{m} := |\widehat{T} \setminus T|$. Consider the set $\mathcal{M} = \{m \in \mathbb{N} : m > s\phi(m \wedge n) \cdot 2(2\bar{c}/\kappa_1 + 3(\bar{c}-1))^2\}$. With probability at least $1-\alpha$*

$$\widehat{m} \leq s \cdot \min_{m \in \mathcal{M}} \phi(m \wedge n) \left(\frac{2\bar{c}}{\kappa_1} + 3(\bar{c}-1) \right)^2 .$$

The main implication of Theorem 5 is that if $\min_{m\in\mathcal{M}} \phi(m \wedge n) \lesssim_P 1$, which we show below to be valid in Lemmas 9 and 10 for important designs, then with probability at least $1-\alpha$

$$\widehat{m} \lesssim_P s. \tag{5.25}$$

Consequently, for these designs, LASSO's sparsity is of the same order as the oracle sparsity, namely $\widehat{s} := |\widehat{T}| \leq s + \widehat{m} \lesssim_P s$ with high probability. The reason for this is that $\min_{m\in\mathcal{M}} \phi(m) \ll \phi(n)$ for these designs, which allows us to sharpen the previous sparsity bound (5.24) considered in [2] and [13]. Also, our new bound is comparable to the bounds in [20] in terms of order of sharpness, but it requires a smaller penalty level λ which also does not depend on the unknown sparse eigenvalues as in [20].

Next we show that $\min_{m\in\mathcal{M}} \phi(m \wedge n) \lesssim_P 1$ for two very common designs of interest, so that the bound (5.25) holds as a consequence. As a side contribution, we also show that for these designs all the restricted sparse eigenvalues and restricted eigenvalues defined earlier behave

nicely. We state these results in asymptotic form for the sake of exposition, although we can convert them to finite sample form using the results in [20] and Lemma 7.

The following lemma deals with a Gaussian design; it uses the standard concept of (unrestricted) sparse eigenvalues (see, e.g. [2]) to state a primitive condition on the population design matrix.

Lemma 9 (Gaussian design). *Suppose $\tilde{x}_i$, $i = 1, \ldots, n$, are i.i.d. zero-mean Gaussian random vectors, such that the population design matrix $\mathrm{E}[\tilde{x}_i\tilde{x}_i']$ has ones on the diagonal, and its $s \log n$-sparse eigenvalues are bounded from above by $\varphi < \infty$ and bounded from below by $\kappa^2 > 0$. Define x_i as a normalized form of $\tilde{x}_i$, namely $x_{ij} = \tilde{x}_{ij}/\sqrt{\mathbb{E}_n[\tilde{x}_{ij}^2]}$. Then for any $m \leq (s\log(n/e)) \wedge (n/[16\log p])$, with probability at least $1 - 2\exp(-n/16)$,*

$$\phi(m) \leq 8\varphi, \quad \widetilde{\kappa}(m)^2 \geq \kappa^2/72, \quad \textit{and} \quad \mu_m \leq 24\sqrt{\varphi}/\kappa.$$

Therefore, under the conditions of Theorem 5 and $n/(s\log p) \to \infty$, we have that as $n \to \infty$

$$\widehat{m} \leq s \cdot (8\varphi)\left(\frac{2\bar{c}}{\kappa_1} + 3(\bar{c}-1)\right)^2$$

with probability approaching at least $1 - \alpha$, where we can take $\kappa_1 \geq \kappa/24$.

The following lemma deals with arbitrary bounded regressors.

Lemma 10 (Bounded design). *Suppose $\tilde{x}_i$ $i = 1, \ldots, n$, are i.i.d. bounded zero-mean random vectors, with $\max_{1\leq i\leq n, 1\leq j\leq p} |\tilde{x}_{ij}| \leq K_B$ for all n and p. Assume that the population design matrix $\mathrm{E}[\tilde{x}_i\tilde{x}_i']$ has ones on the diagonal, and its $s\log n$-sparse eigenvalues are bounded from above by $\varphi < \infty$ and bounded from below by $\kappa^2 > 0$. Define x_i as a normalized form of $\tilde{x}_i$, namely $x_{ij} = \tilde{x}_{ij}/\sqrt{\mathbb{E}_n[\tilde{x}_{ij}^2]}$. Then there is a constant $\epsilon > 0$ such that if $\sqrt{n}/K_B \to \infty$ and $m \leq (s\log(n/e)) \wedge ([\epsilon/K_B]\sqrt{n/\log p})$, we have that as $n \to \infty$*

$$\phi(m) \leq 4\varphi, \quad \widetilde{\kappa}(m)^2 \geq \kappa^2/4, \quad \textit{and} \quad \mu_m \leq 4\sqrt{\varphi}/\kappa,$$

with probability approaching 1. Therefore, under the conditions of Theorem 5 and provided $\sqrt{n}/(K_B s\sqrt{\log p}) \to \infty$, we have that as $n \to \infty$,

$$\widehat{m} \leq s \cdot (4\varphi)\left(\frac{2\bar{c}}{\kappa_1} + 3(\bar{c}-1)\right)^2$$

with probability approaching at least $1 - \alpha$, where we can take $\kappa_1 \geq \kappa/8$.

Proof of Theorem 5. The choice of λ implies that with probability at least $1-\alpha$ we have $\lambda \geq c \cdot n\|S\|_\infty$. In that event, by Lemma 7

$$\sqrt{\widehat{m}} \leq \sqrt{\phi(\widehat{m})} \cdot 2\bar{c}\sqrt{s}/\kappa_1 + 3(\bar{c}+1)\sqrt{\phi(\widehat{m})} \cdot nc_s/\lambda,$$

which can be rewritten as

$$\widehat{m} \leq s \cdot \phi(\widehat{m}) \left(\frac{2\bar{c}}{\kappa_1} + 3(\bar{c}+1)\frac{nc_s}{\lambda\sqrt{s}} \right)^2. \tag{5.26}$$

Note that $\widehat{m} \leq n$ by optimality conditions. Consider any $M \in \mathcal{M}$, and suppose $\widehat{m} > M$. Therefore by Lemma 8 on sublinearity of sparse eigenvalues

$$\widehat{m} \leq s \cdot \left\lceil \frac{\widehat{m}}{M} \right\rceil \phi(M) \left(\frac{2\bar{c}}{\kappa_1} + 3(\bar{c}+1)\frac{nc_s}{\lambda\sqrt{s}} \right)^2.$$

Thus, since $\lceil k \rceil \leq 2k$ for any $k \geq 1$ we have

$$M \leq s \cdot 2\phi(M) \left(\frac{2\bar{c}}{\kappa_1} + 3(\bar{c}+1)\frac{nc_s}{\lambda\sqrt{s}} \right)^2$$

which violates the condition on M and s since $c_s \leq \sigma\sqrt{s/n}$, $\lambda \geq c\sigma\sqrt{n}$, and $(\bar{c}+1)/c = \bar{c}-1$. Therefore, we must have $\widehat{m} \leq M$.

In turn, applying (5.26) once more with $\widehat{m} \leq (M \wedge n)$ we obtain

$$\widehat{m} \leq s \cdot \phi(M \wedge n) \left(\frac{2\bar{c}}{\kappa_1} + 3(\bar{c}+1)\frac{nc_s}{\lambda\sqrt{s}} \right)^2.$$

Further, using again that $c_s \leq \sigma\sqrt{s/n}$ and $\lambda \geq c\sigma\sqrt{n}$ we have

$$\widehat{m} \leq s \cdot \phi(M \wedge n) \left(\frac{2\bar{c}}{\kappa_1} + 3(\bar{c}-1) \right)^2,$$

since $(\bar{c}+1)/c = \bar{c}-1$. The result follows by minimizing the bound over $M \in \mathcal{M}$. □

5.2. **Performance of the post-LASSO Estimator.** Here we show that the post-LASSO estimator enjoys good theoretical performance despite possibly "poor" selection of the model by LASSO.

Theorem 6 (Performance of post-LASSO). *In either the parametric model or the nonparametric model, if $\lambda \geq cn\|S\|_\infty$, for any $\varepsilon > 0$ there is a constant K_ε independent of n such that with probability at least $1-\varepsilon$*

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \leq K_\varepsilon \sigma \sqrt{\frac{\widehat{m}\log p + (\widehat{m}+s)\log \mu_{\widehat{m}}}{n}} + 2c_s + 1\{T \not\subseteq \widehat{T}\}\sqrt{\frac{\lambda\sqrt{s}}{n\kappa_1} \cdot \left(\frac{(1+c)\lambda\sqrt{s}}{cn\kappa_1} + 2c_s \right)},$$

where $\widehat{m} := |\widehat{T} \setminus T|$ *and* $c_s = 0$ *in the parametric model. In particular, under the data-driven choice of* λ *specified in (2.3) with* $\log(1/\alpha) \lesssim \log p$*, for any* $\varepsilon > 0$ *there is a constant* $K'_{\varepsilon,\alpha}$ *such that*

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \leq K'_{\varepsilon,\alpha}\sigma \left[\sqrt{\frac{\widehat{m}\log(p\mu_{\widehat{m}})}{n}} + \sqrt{\frac{s\log\mu_{\widehat{m}}}{n}} + 1\{T \not\subseteq \widehat{T}\}\sqrt{\frac{s\log p}{n}\frac{1}{\kappa_1}} \right] \tag{5.27}$$

with probability at least $1 - \alpha - \varepsilon$.

Proof of Theorem 6. Note that by the optimality of $\widehat{\beta}$ in the LASSO problem, and letting $\widehat{\delta} = \widehat{\beta} - \beta_0$,

$$B_n := \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) \ \leq \tfrac{\lambda}{n}(\|\beta_0\|_1 - \|\widehat{\beta}\|_1) \leq \tfrac{\lambda}{n}(\|\widehat{\delta}_T\|_1 - \|\widehat{\delta}_{T^c}\|_1). \tag{5.28}$$

If $B_n := \|\widehat{\delta}_{T^c}\|_1 > \bar{c}\|\widehat{\delta}_T\|_1$, we have $\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) \leq 0$ since $\bar{c} \geq 1$. Otherwise, if $\|\widehat{\delta}_{T^c}\|_1 \leq \bar{c}\|\widehat{\delta}_T\|_1$, by RE.1($c$) we have

$$B_n := \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) \leq \frac{\lambda}{n}\|\widehat{\delta}_T\|_1 \leq \frac{\lambda}{n}\frac{\sqrt{s}\|\widehat{\delta}\|_{2,n}}{\kappa_1}. \tag{5.29}$$

The result follows by applying Lemma 2 to bound $\|\widehat{\delta}\|_{2,n}$ and Theorem 1, and also noting that if $T \subseteq \widehat{T}$ we have $C_n = 0$ so that $B_n \wedge C_n \leq 1\{T \not\subseteq \widehat{T}\}B_n$.

The second claim is immediate from the first, using the condition that $c_s \lesssim \sigma\sqrt{s/n}$, relation (2.9), in the case of the nonparametric model. □

This theorem provides a performance bound for post-LASSO as a function of 1) LASSO's sparsity characterized by $\widehat{m}$, 2) LASSO's rate of convergence, and 3) LASSO's model selection ability. For common designs this bound implies that post-LASSO performs at least as well as LASSO, but it can be strictly better in some cases, and has smaller regularization bias. We provide further theoretical comparisons in what follows, and computational examples supporting these comparisons appear in Section 6. It is also worth repeating here that performance bounds in other norms of interest immediately follow by the triangle inequality and by definition of $\widetilde{\kappa}$:

$$\sqrt{\mathbb{E}_n[x_i'\widetilde{\beta} - f_i]^2} \leq \|\widetilde{\beta} - \beta_0\|_{2,n} + c_s \ \text{ and } \|\widetilde{\beta} - \beta_0\|_2 \leq \|\widetilde{\beta} - \beta_0\|_{2,n}/\widetilde{\kappa}(\widehat{m}). \tag{5.30}$$

Comment 5.1 (Comparison of the performance of post-LASSO vs LASSO). In order to carry out complete and formal comparisons between LASSO and post-LASSO, we assume that

$$\phi(\widehat{m}) \lesssim_P 1, \ \ \kappa_1 \gtrsim_P 1, \ \mu_{\widehat{m}} \lesssim_P 1, \ \ \log(1/\alpha) \lesssim \log p \ \text{ and } \ \alpha = o(1). \tag{5.31}$$

We established fairly general sufficient conditions for the first three relations in Lemmas 9 and 10. The fourth relation is a mild condition on the choice of α in the definition of the data-driven

choice (2.3) of penalty level λ, which simplifies the probability statements in what follows. We first note that under (5.31) post-LASSO with the data-driven penalty level λ specified in (2.3) obeys:

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \lesssim_P \sigma \left[\sqrt{\frac{\widehat{m}\log p}{n}} + \sqrt{\frac{s}{n}} + 1\{T \not\subseteq \widehat{T}\}\sqrt{\frac{s\log p}{n}} \right].$$

In addition, conditions (5.31) and Theorem 5 imply the oracle sparsity $\widehat{m} \lesssim_P s$.

It follows that post-LASSO generally achieves the same near-oracle rate as LASSO:

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \lesssim_P \sigma\sqrt{\frac{s\log p}{n}}. \tag{5.32}$$

Notably, this occurs despite the fact that LASSO may in general fail to correctly select the oracle model T as a subset, that is $T \not\subseteq \widehat{T}$.

Furthermore, there is a class of well-behaved models – a neighborhood of parametric models with well-separated coefficients – in which post-LASSO strictly improves upon LASSO. Specifically, if $\widehat{m} = o_P(s)$ and $T \subseteq \widehat{T}$ wp $\to 1$, as under conditions of Lemmas 3 and 4, then

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \lesssim_P \sigma \left[\sqrt{\frac{o(s)\log p}{n}} + \sqrt{\frac{s}{n}} \right]. \tag{5.33}$$

That is, post-LASSO strictly improves upon LASSO's rate. Finally, in the extreme case of perfect model selection, when $\widehat{m} = 0$ and $T \subseteq \widehat{T}$ wp $\to 1$, as under conditions of Lemma 4, post-LASSO naturally achieves the oracle performance: $\|\widetilde{\beta} - \beta_0\|_{2,n} \lesssim_P \sigma\sqrt{s/n}$. $\square$

5.3. **Performance of the post-goof-trimmed LASSO estimator.** In what follows we provide performance bounds for the post-goof-trimmed estimator $\widetilde{\beta}$ defined in equation (4.20) for the case where the first-step estimator $\widehat{\beta}$ is LASSO.

Theorem 7 (Performance of post-goof-trimmed LASSO). *In either the parametric model or the nonparametric model, if $\lambda \geq cn\|S\|_\infty$, for any $\varepsilon > 0$ there is a constant K_ε independent of n such that with probability at least $1 - \varepsilon$*

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \leq K_\varepsilon \sigma \sqrt{\frac{\widetilde{m}\log p + (\widetilde{m} + s)\log \mu_{\widetilde{m}}}{n}} + 2c_s + 1\{T \not\subseteq \widetilde{T}\}\sqrt{\frac{\lambda\sqrt{s}}{n\kappa_1}\left(\frac{(1+c)\lambda\sqrt{s}}{cn\kappa_1} + 2c_s\right)},$$

where $\widetilde{m} := |\widetilde{T} \setminus T|$ and $c_s = 0$ in the parametric case. Under the data-driven choice of λ specified in (2.3) with $\log(1/\alpha) \lesssim \log p$, for any $\varepsilon > 0$ there is a constant $K'_{\varepsilon,\alpha}$ such that

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \leq K'_{\varepsilon,\alpha}\sigma \left[\sqrt{\frac{\widetilde{m}\log(p\mu_{\widetilde{m}})}{n}} + \sqrt{\frac{s\log \mu_{\widetilde{m}}}{n}} + 1\{T \not\subseteq \widetilde{T}\}\sqrt{\frac{s\log p}{n}\frac{1}{\kappa_1}} \right] \tag{5.34}$$

with probability at least $1 - \alpha - \varepsilon$.

Proof. The proof of the first claim follows the same steps as the proof of Theorem 6, invoking Theorem 3 in the last step. The second claim follows immediately from the first, where we also use the condition $c_s \lesssim \sigma\sqrt{s/n}$ from (2.9) in the nonparametric model, in addition the condition $\gamma \leq 0$ imposed in the construction of the estimator. □

This theorem provides a performance bound for post-goof-trimmed LASSO as a function of 1) its sparsity characterized by $\widetilde{m}$, 2) LASSO's rate of convergence, and 3) the model selection ability of the trimming scheme. Generally, this bound is at least as good as the bound for post-LASSO, since the post-goof-trimmed LASSO trims as much as possible subject to maintaining certain goodness-of-fit. It is also appealing that this estimator determines the trimming level in a completely data-driven fashion. Moreover, by construction the estimated model is sparser than post-LASSO's model, which leads to the superior performance of post-goof-trimmed LASSO over post-LASSO in some cases. We further provide further theoretical comparisons below and computational examples in Section 6.

Comment 5.2 (Comparison of the performance of post-goof-trimmed LASSO vs LASSO and post-LASSO). In order to carry out complete and formal comparisons, we assume condition (5.31) as before. Under these conditions, post-goof-trimmed LASSO obeys the following performance bound:

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \lesssim_P \sigma \left[\sqrt{\frac{\widetilde{m}\log p}{n}} + \sqrt{\frac{s}{n}} + 1\{T \not\subseteq \widetilde{T}\}\sqrt{\frac{s\log p}{n}} \right],$$

which is analogous to the bound for post-LASSO, since $\widetilde{m} \leq \widehat{m} \lesssim_P s$ by conditions (5.31) and Theorem 5. It follows that in general post-goof-trimmed LASSO matches the near oracle rate of convergence of LASSO and post-LASSO:

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \lesssim_P \sigma\sqrt{\frac{s\log p}{n}}. \tag{5.35}$$

Nonetheless, there is a class of models – a neighborhood of parametric models with well-separated coefficients – for which improvements upon the rate of convergence of LASSO is possible. Specifically, if $\widetilde{m} = o_P(s)$ and $T \subseteq \widetilde{T}$ wp $\to 1$ then we obtain the performance bound (5.33), that is, post-goof-trimmed LASSO strictly improves upon LASSO's rate. Furthermore, if $\widetilde{m} = o_P(\widehat{m})$ and $T \subseteq \widetilde{T}$ wp $\to 1$, post-goof-trimmed LASSO also outperforms post-LASSO:

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \lesssim_P \sigma \left[\sqrt{\frac{o(\widehat{m})\log p}{n}} + \sqrt{\frac{s}{n}} \right].$$

Lastly, under conditions of Lemma 3 holding for $t = t_\gamma$, post-goof-trimmed LASSO achieves the oracle performance, $\|\widetilde{\beta} - \beta_0\|_{2,n} \lesssim_P \sigma\sqrt{s/n}$. □

5.4. **Performance of the post-traditional-trimmed LASSO estimator.** Next we consider the traditional trimming scheme which truncates to zero all components below a set threshold t. This is arguably the most used trimming scheme in the literature. To state the result, recall that $\widehat{\beta}_{tj} = \widehat{\beta}_j 1\{|\widehat{\beta}_j| > t\}$, $\widetilde{m} := |\widetilde{T} \setminus T|$, $m_t := |\widehat{T} \setminus \widetilde{T}|$ and $\gamma_t := \|\widehat{\beta}_t - \widehat{\beta}\|_{2,n}$ where $\widehat{\beta}$ is the LASSO estimator.

Theorem 8 (Performance of post-traditional-trimmed LASSO). *In either the parametric model or the nonparametric model, if $\lambda \geq cn\|S\|_\infty$, for any $\varepsilon > 0$ there is a constant K_ε independent of n such that with probability at least $1 - \varepsilon$ we have*

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \leq K_\varepsilon \sigma \sqrt{\frac{\widetilde{m}\log p + (\widetilde{m} + s)\log \mu_{\widetilde{m}}}{n}} + 2c_s +$$
$$+1\{T \not\subseteq \widetilde{T}\} \sqrt{\gamma_t(K_\varepsilon \sigma G_t + 6c_s + \gamma_t) + \frac{\lambda\sqrt{s}}{n\kappa_1}\left(\frac{2\gamma_t(1+c)}{c} + \frac{(1+c)\lambda\sqrt{s}}{cn\kappa_1} + 2c_s\right)},$$

where $G_t = \sqrt{m_t \log(p\mu_{m_t})}/\sqrt{n}$ and $\gamma_t \leq t\sqrt{\phi(m_t)m_t}$. Under the data-driven choice of λ specified in (2.3) for $\log(1/\alpha) \lesssim \log p$, for any $\varepsilon > 0$ there is a constant $K'_{\varepsilon,\alpha}$ such that with probability at least $1 - \alpha - \varepsilon$

$$\begin{aligned}\|\widetilde{\beta} - \beta_0\|_{2,n} \quad \leq \quad & K'_{\varepsilon,\alpha}\left[\sigma\sqrt{\frac{\widetilde{m}\log(p\mu_{\widetilde{m}})}{n}} + \sigma\sqrt{\frac{s\log\mu_{\widetilde{m}}}{n}} + \right.\\ & \left. +1\{T \not\subseteq \widetilde{T}\}\left(\gamma_t + \sqrt{\gamma_t\sigma\sqrt{\frac{m_t\log(p\mu_{m_t})}{n}}} + \sigma\sqrt{\frac{s\log p}{n}\frac{1}{\kappa_1}}\right)\right].\end{aligned}$$

Proof. The proof of the first claim follows the same steps as the proof of Theorem 6; invoking Theorem 4 in the last step. The second claim follows from the first, where we also use the condition $c_s \lesssim \sigma\sqrt{s/n}$, relation (2.9), for the nonparametric model. □

This theorem provides a performance bound for post-traditional-trimmed LASSO as a function of 1) its sparsity characterized by $\widetilde{m}$ and improvements in sparsity over LASSO characterized by m_t, 2) LASSO's rate of convergence, 3) the trimming threshold t and resulting goodness-of-fit loss γ_t relative to LASSO induced by trimming, and 4) model selection ability of the trimming scheme. Generally, this bound may be worse than the bound for LASSO, and

this arises because the post-traditional-trimmed LASSO may potentially use too much trimming resulting in large goodness-of-fit losses γ_t. We provide further theoretical comparisons below and computational examples in Section 6.

Comment 5.3 (Comparison of the performance of post-traditional-trimmed LASSO vs LASSO and post-LASSO). In this discussion we also assume conditions (5.31) made in the previous formal comparisons. Under these conditions, post-traditional-trimmed LASSO obeys the bound:

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \leq \sigma\sqrt{\frac{\widetilde{m}\log p}{n}} + \sigma\sqrt{\frac{s}{n}} + 1\{T \not\subseteq \widetilde{T}\}\left(\gamma_t \vee \sigma\sqrt{\frac{s\log p}{n}}\right). \tag{5.36}$$

In this case we have $\widetilde{m} \vee m_t \leq s + \widehat{m} \lesssim_P s$ by Theorem 5, and, in general, the rate above cannot improve upon LASSO's rate of convergence given in Lemma 2.

As expected, the choice of t, which controls γ_t via the the bound $\gamma_t \leq t\sqrt{\phi(m_t)m_t}$, can have a large impact on the performance bounds:

$$t \lesssim \sigma\sqrt{\frac{\log p}{n}} \quad \Longrightarrow \quad \|\widetilde{\beta} - \beta_0\|_{2,n} \lesssim_P \sigma\sqrt{\frac{s\log p}{n}} \tag{5.37}$$

$$t \lesssim \sigma\sqrt{\frac{s\log p}{n}} \quad \Longrightarrow \quad \|\widetilde{\beta} - \beta_0\|_{2,n} \lesssim_P \sigma\sqrt{\frac{s^2\log p}{n}}. \tag{5.38}$$

Both options are standard suggestions in the literature on model selection via LASSO, as we reviewed in Lemma 3 parts (2) and (3). The first choice (5.37), suggested by [11], is theoretically sound, since it guarantees that post-traditional-trimmed LASSO achieves the near-oracle rate of LASSO. The second choice, however, results in a very poor performance bound, and even suggests inconsistency if s^2 is large relative to n. Note that to implement the first choice (5.37) in practice we can set $t = \lambda/n$.

Furthermore, there is a special class of models – a neighborhood of parametric models with well-separated coefficients – for which improvements upon the rate of convergence of LASSO is possible. Specifically, if $\widetilde{m} = o_P(s)$ and $T \subseteq \widetilde{T}$ wp $\to 1$ then we recover the performance bound (5.33), that is, post-traditional-trimmed LASSO strictly improves upon LASSO's rate. Furthermore, if $\widetilde{m} = o_P(\widehat{m})$ and $T \subseteq \widetilde{T}$ wp $\to 1$, post-traditional-trimmed LASSO also outperforms post-LASSO:

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \lesssim_P \quad \sigma\left[\sqrt{\frac{o(\widehat{m})\log p}{n}} + \sqrt{\frac{s}{n}}\right].$$

Lastly, under the conditions of Lemma 3 holding for the given t, post-traditional-trimmed LASSO achieves the oracle performance, $\|\widetilde{\beta} - \beta_0\|_{2,n} \lesssim_P \sigma\sqrt{s/n}$. □

6. Empirical Performance Relative to LASSO

In this section we assess the finite sample performance of the following estimators: 1) LASSO, which is our benchmark, 2) post-LASSO, 3) post-goof-trimmed LASSO, and 4) post-traditional-trimmed LASSO with the trimming threshold $t = \lambda/n$ suggested by Lemma 3 part (3). We consider a "parametric" and a "nonparametric" model of the form:

$$y_i = f_i + \epsilon_i, \quad f_i = x_i'\theta_0, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, ..., n,$$

where in the parametric model

$$\theta_0 = C \cdot [1, 1, 1, 1, 1, 0, 0, ..., 0]', \tag{6.39}$$

and in the nonparametric model

$$\theta_0 = C \cdot [1, 1/2, 1/3, ..., 1/p]'. \tag{6.40}$$

The parameter C determines the size of the coefficients, representing the "strength of the signal", and we vary C between 0 and 2. The number of regressors is $p = 500$, the sample size is $n = 100$, the variance of the noise is $\sigma^2 = 1$, and we used 1000 simulations for each design. We generate regressors from the normal law $x_i \sim N(0, \Sigma)$, and consider three designs of the covariance matrix Σ: a) the isotropic design with $\Sigma_{jk} = 0$ for $j \neq k$, b) the Toeplitz design with $\Sigma_{jk} = (1/2)^{|j-k|}$, and c) the equi-correlated design with $\Sigma_{jk} = 1/2$ for $j \neq k$; in all designs $\Sigma_{jj} = 1$. Thus our parametric model is very sparse and offers a rather favorable setting for applying LASSO-type methods, while our nonparametric model is non-sparse and much less favorable.

We present the results of computational experiments for each design a)-c) in Figures 2-4. The left column of each figure reports the results for the parametric model, and the right column of each figure reports the results for the nonparametric model. For each model the figures plot the following as a function of the signal strength for each estimator $\widetilde{\beta}$:

- in the top panel, the number of regressors selected, $|\widetilde{T}|$,
- in the middle panel, the norm of the bias, namely $\|E[\widetilde{\beta} - \theta_0]\|$,
- in the bottom panel, the average empirical risk, namely $E[\mathbb{E}_n[f_i - x_i'\widetilde{\beta}]^2]$.

We will focus the discussion on the isotropic design, and only highlight differences for other designs.

Figure 2, left panel, shows the results for the parametric model with the isotropic design. We see from the bottom panel that, for a wide range of signal strength C, both post-LASSO and

post-goof-trimmed LASSO significantly outperform both LASSO and post-traditional-trimmed LASSO in terms of empirical risk. The middle panel shows that the first two estimators' superior performance stems from their much smaller bias. We see from the top panel that LASSO achieves good sparsity, ensuring that post-LASSO performs well, but post-goof-trimmed LASSO achieves even better sparsity. Under very high signal strength, post-goof-trimmed LASSO achieves the performance of the oracle estimator; post-traditional-trimmed LASSO also achieves this performance; post-LASSO nearly matches it; while LASSO does not match this performance. Interestingly, the post-traditional-trimmed LASSO performs very poorly for intermediate ranges of signal.

Figure 2, right panel, shows the results for the nonparametric model with the isotropic design. We see from the bottom panel that, as in the parametric model, both post-LASSO and post-goof-trimmed LASSO significantly outperform both LASSO and post-traditional-trimmed LASSO in terms of empirical risk. As in the parametric model, the middle panel shows that the first two estimators are able to outperform the last two because they have a much smaller bias. We also see from the top panel that, as in the parametric model, LASSO achieves good sparsity, while post-goof-trimmed LASSO achieves excellent sparsity. In contrast to the parametric model, in the nonparametric setting the post-traditional-trimmed LASSO performs poorly in terms of empirical risk for almost all signals, except for very weak signals. Also in contrast to the parametric model, no estimator achieves the exact oracle performance, although LASSO, and especially post-LASSO and post-goof-trimmed LASSO perform nearly as well, as we would expect from the theoretical results.

Figure 3 shows the results for the parametric and nonparametric model with the Toeplitz design. This design deviates only moderately from the isotropic design, and we see that all of the previous findings continue to hold. Figure 4 shows the results under the equi-correlated design. This design strongly deviates from the isotropic design, but we still see that the previous findings continue to hold with only a few differences. Specifically, we see from the top panels that in this case LASSO no longer selects very sparse models, while post-goof-trimmed LASSO continues to perform well and selects very sparse models. Consequently, in the case of the parametric model, post-goof-trimmed LASSO substantially outperforms post-LASSO in terms of empirical risk, as the bottom-left panel shows. In contrast, we see from the bottom right panel that in the nonparametric model, post-goof-trimmed LASSO performs equally as well as post-LASSO in terms of empirical risk, despite the fact that it uses a much sparser model for estimation.

The findings above confirm our theoretical results on post-penalized estimators in parametric and nonparametric models. Indeed, we see that post-goof-trimmed LASSO and post-LASSO are at least as good as LASSO, and often perform considerably better since they remove penalization bias. Post-goof-trimmed LASSO outperforms post-LASSO whenever LASSO does not produce excellent sparsity. Moreover, when the signal is strong and the model is parametric and sparse (or very close to being such), the LASSO-based model selection permits the selection of oracle or near-oracle model. That allows for post-model selection estimators to achieve improvements in empirical risk over LASSO. Of particular note is the excellent performance of post-goof-trimmed LASSO, which uses data-driven trimming to select a sparse model. This performance is fully consistent with our theoretical results. Finally, traditional trimming performs poorly for intermediate ranges of signal. In particular, it exhibits very large biases leading to large goodness-of-fit losses.

Appendix A. Proofs of Lemmas 1 and 2

Proof of Lemma 1. Following Bickel, Ritov and Tsybakov [2], to establish the result we make the use of the following relations for $\delta = \widehat{\beta} - \beta$ and for $\lambda \geq cn\|S\|_\infty$:

$$\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) \geq -\|S\|_\infty \|\delta\|_1 + \|\delta\|_{2,n}^2 \geq -\frac{\lambda}{cn}(\|\delta_T\|_1 + \|\delta_{T^c}\|_1) + \|\delta\|_{2,n}^2 \qquad \text{(A.41)}$$

$$\|\beta_0\|_1 - \|\widehat{\beta}\|_1 = \|\beta_{0T}\|_1 - \|\widehat{\beta}_T\|_1 - \|\widehat{\beta}_{T^c}\|_1 \leq \|\delta_T\|_1 - \|\delta_{T^c}\|_1. \qquad \text{(A.42)}$$

By definition of $\widehat{\beta}$, $\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) \leq \frac{\lambda}{n}\|\beta_0\|_1 - \frac{\lambda}{n}\|\widehat{\beta}\|_1$, which, by (A.41) and (A.42), implies that

$$-\frac{\lambda}{cn}(\|\delta_T\|_1 + \|\delta_{T^c}\|_1) + \|\delta\|_{2,n}^2 \leq \frac{\lambda}{n}(\|\delta_T\|_1 - \|\delta_{T^c}\|_1). \qquad \text{(A.43)}$$

Since $\|\delta\|_{2,n}^2 \geq 0$,

$$\|\delta_{T^c}\|_1 \leq \frac{c+1}{c-1} \cdot \|\delta_T\|_1 = \bar{c}\|\delta_T\|_1. \qquad \text{(A.44)}$$

Going back to (A.43), we get that:

$$\|\delta\|_{2,n}^2 \leq \left(1 + \frac{1}{c}\right)\frac{\lambda}{n}\|\delta_T\|_1 \leq \left(1 + \frac{1}{c}\right)\frac{\lambda}{n}\sqrt{s}\frac{\|\delta\|_{2,n}}{\kappa_1}$$

where we used that $c \geq 1$ and invoked RE.1(c) since (A.44) holds. Solve for $\|\delta\|_{2,n}$.

Finally, the bound on $\Lambda(1-\alpha|X)$ follows from the union bound and a probability inequality for Gaussian random variables, $P(|\xi| > M) \leq \exp(-M^2/2)$ if $\xi \sim N(0,1)$, see Proposition 2.2.1(a) in [8]. □

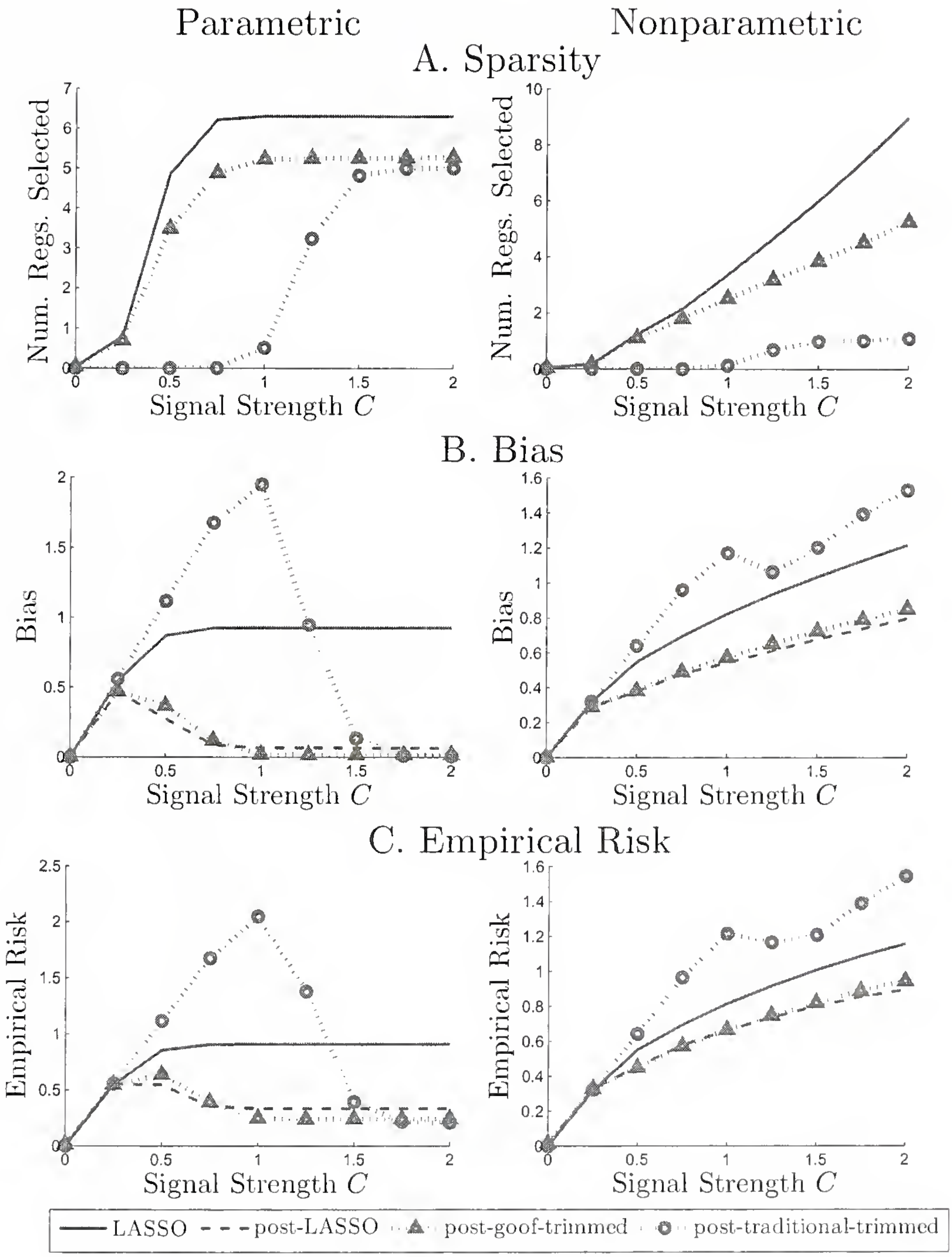

FIGURE 2. This figure plots the performance of the estimators listed in the text under the isotropic design for the covariates, $\Sigma_{jk} = 0$ if $j \neq k$. The left column corresponds to the parametric case and the right column corresponds to the nonparametric case described in the text. The number of regressors is $p = 500$ and the sample size is $n = 100$ with 1000 simulations for each value of C.

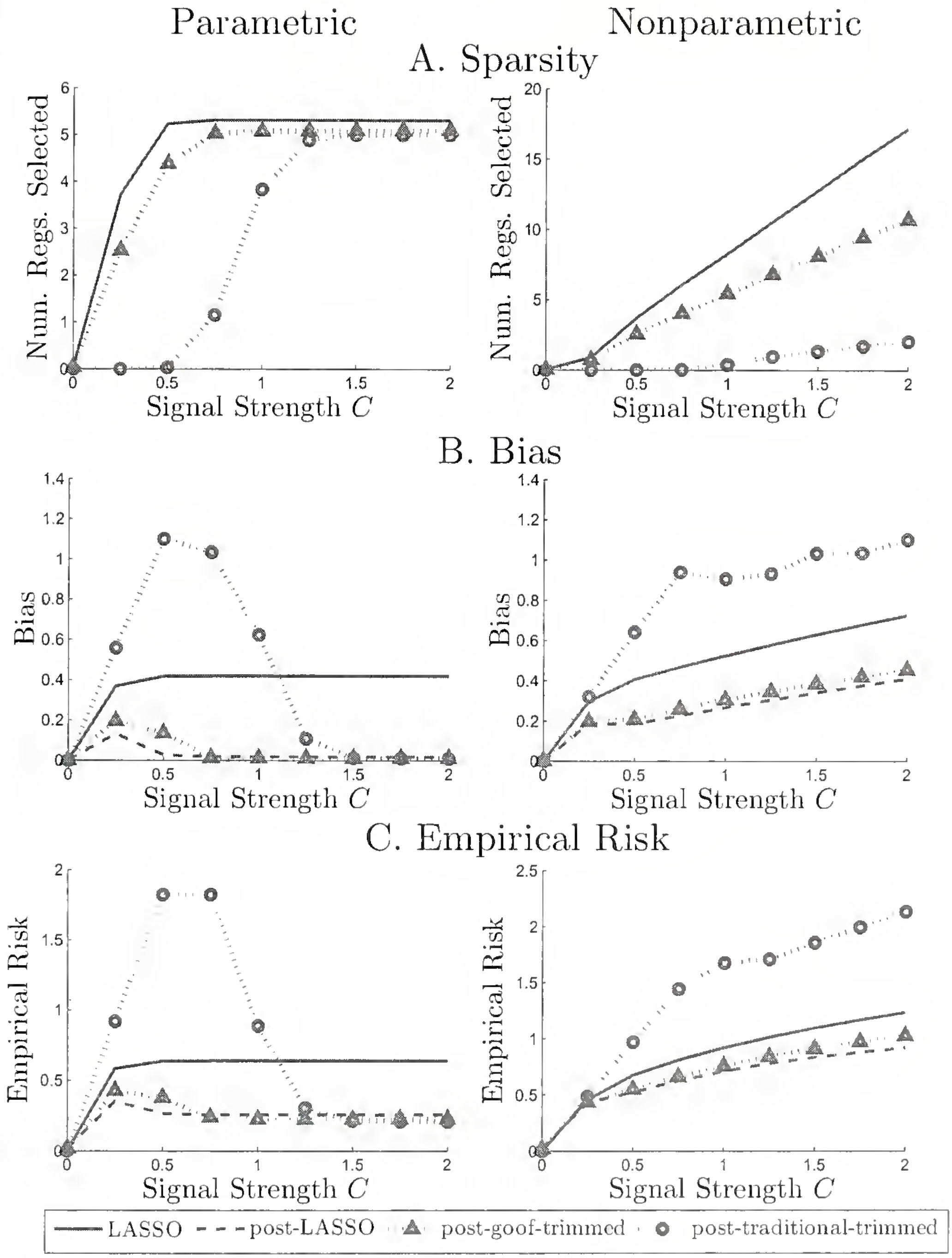

FIGURE 3. This figure plots the performance of the estimators listed in the text under the Toeplitz design for the covariates, $\Sigma_{jk} = \rho^{|j-k|}$ if $j \neq k$. The left column corresponds to the parametric case and the right column corresponds to the nonparametric case described in the text. The number of regressors is $p = 500$ and the sample size is $n = 100$ with 1000 simulations for each value of C.

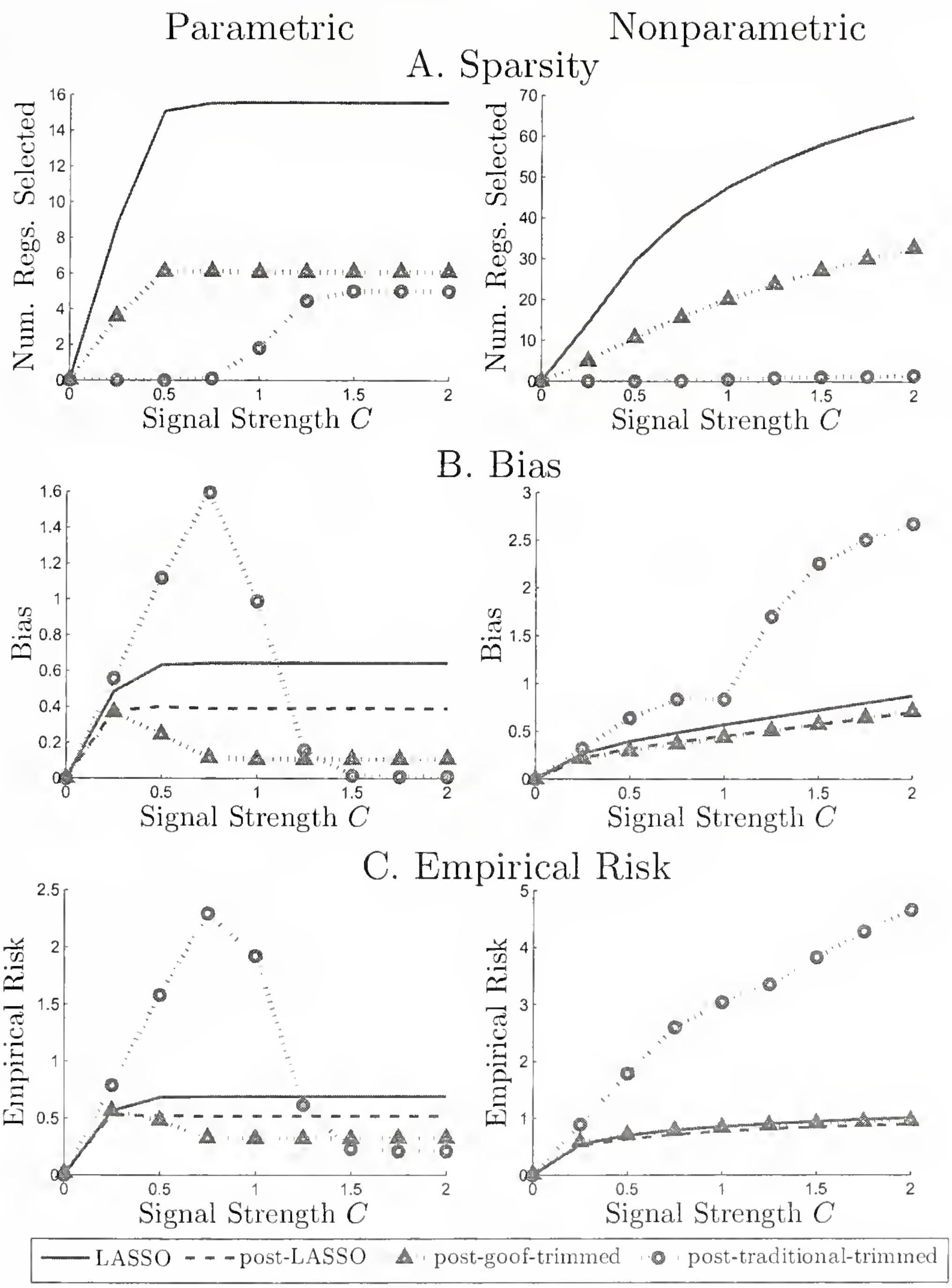

FIGURE 4. This figure plots the performance of the estimators listed in the text under the equi-correlated design for the covariates, $\Sigma_{jk} = \rho$ if $j \neq k$. The left column corresponds to the parametric case and the right column corresponds to the nonparametric case described in the text. The number of regressors is $p = 500$ and the sample size is $n = 100$ with 1000 simulations for each value of C.

Proof of Lemma 2. Similar to [2], to prove Lemma 2 we make the use of the following relation: for $\delta = \widehat{\beta} - \beta_0$, if $\lambda \geq cn\|S\|_\infty$

$$\begin{aligned} \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) - \|\delta\|_{2,n}^2 &= 2\mathbb{E}_n[\epsilon_i x_i'\delta] + 2\mathbb{E}_n[r_i x_i'\delta] \geq -\|S\|_\infty\|\delta\|_1 - 2c_s\|\delta\|_{2,n} \\ &\geq -\frac{\lambda}{cn}(\|\delta_T\|_1 + \|\delta_{T^c}\|_1) - 2c_s\|\delta\|_{2,n} \end{aligned} \tag{A.45}$$

By definition of $\widehat{\beta}$, $\widehat{Q}(\widehat{\beta}) - Q(\beta_0) \leq \frac{\lambda}{n}\|\beta_0\|_1 - \frac{\lambda}{n}\|\widehat{\beta}\|_1$, which implies that

$$-\frac{\lambda}{cn}(\|\delta_T\|_1 + \|\delta_{T^c}\|_1) + \|\delta\|_{2,n}^2 - 2c_s\|\delta\|_{2,n} \leq \frac{\lambda}{n}(\|\delta_T\|_1 - \|\delta_{T^c}\|_1) \tag{A.46}$$

If $\|\delta\|_{2,n}^2 - 2c_s\|\delta\|_{2,n} < 0$, then we have established the bound in the statement of the theorem. On the other hand, if $\|\delta\|_{2,n}^2 - 2c_s\|\delta\|_{2,n} \geq 0$ we get

$$\|\delta_{T^c}\|_1 \leq \frac{c+1}{c-1} \cdot \|\delta_T\|_1 = \bar{c}\|\delta_T\|_1, \tag{A.47}$$

and therefore δ satisfies the domination condition (2.6). From (A.46) and using RE.1(c) we get

$$\|\delta\|_{2,n}^2 - 2c_s\|\delta\|_{2,n} \leq \left(1 + \frac{1}{c}\right)\frac{\lambda}{n}\|\delta_T\|_1 \leq \left(1 + \frac{1}{c}\right)\frac{\sqrt{s}\lambda}{n}\frac{\|\delta\|_{2,n}}{\kappa_1}$$

which gives the result on the prediction norm. Finally, the bound on $\Lambda(1-\alpha|X)$ follows from the union bound and a probability inequality for Gaussian random variables, $P(|\xi| > M) \leq \exp(-M^2/2)$ if $\xi \sim N(0,1)$, see Proposition 2.2.1(a) in [8]. □

Appendix B. Proofs of Lemmas for Post-Model Selection Estimators

Proof of Lemma 5. The bound on B_n follows from Lemma 6 result (1). The bound on C_n follows from Lemma 6 result (2). □

Proof of Lemma 6. Result (1) follows from the relation

$$|\widehat{Q}(\beta_0 + \delta) - \widehat{Q}(\beta_0) - \|\delta\|_{2,n}^2| = |2\mathbb{E}_n[\epsilon_i x_i'\delta] + 2\mathbb{E}_n[r_i x_i'\delta]|,$$

then applying Theorem 2 on sparse control of noise to $|2\mathbb{E}_n[\epsilon_i x_i'\delta]|$, bounding $|2\mathbb{E}_n[r_i x_i'\delta]|$ by $2c_s\|\delta\|_{2,n}$ using the Cauchy-Schwartz inequality, and bounding $\binom{p}{m}$ by p^m.

Result (2) also follows from Theorem 2 but applying it with $s = 0$, $p = s$ (since only the components in T are modified), $m = k$, and noting that we can take μ_m with $m = 0$. □

Appendix C. Proofs of Lemmas for Sparsity of the LASSO estimator

Proof of Lemma 7. Let $\widehat{T} = \text{support}(\widehat{\beta})$, and $\widehat{m} = |\widehat{T} \setminus T|$. We have from the optimality conditions that

$$2\mathbb{E}_n[x_{ij}(y_i - x_i'\widehat{\beta})] = \text{sign}(\widehat{\beta}_j)\lambda/n \quad \text{for each} \quad j \in \widehat{T} \setminus T.$$

Therefore we have for $R = (r_1, \ldots, r_n)'$

$$\begin{aligned}
\sqrt{\widehat{m}}\lambda &= 2\|(X'(Y - X\widehat{\beta}))_{\widehat{T}\setminus T}\| \\
&\leq 2\|(X'(Y - R - X\beta_0))_{\widehat{T}\setminus T}\| + 2\|(X'R)_{\widehat{T}\setminus T}\| + 2\|(X'X(\beta_0 - \widehat{\beta}))_{\widehat{T}\setminus T}\| \\
&\leq \sqrt{\widehat{m}} \cdot n\|S\|_\infty + 2n\sqrt{\phi(\widehat{m})}c_s + 2n\sqrt{\phi(\widehat{m})}\|\widehat{\beta} - \beta_0\|_{2,n},
\end{aligned}$$

where we used that

$$\begin{aligned}
\|(X'X(\beta_0 - \widehat{\beta}))_{\widehat{T}\setminus T}\| &= \sup_{\|\alpha\|_0 \leq \widehat{m}, \|\alpha\| \leq 1} |\alpha' X'X(\beta_0 - \widehat{\beta})| \\
&\leq \sup_{\|\alpha\|_0 \leq \widehat{m}, \|\alpha\| \leq 1} \|\alpha' X'\| \|X(\beta_0 - \widehat{\beta})\| \\
&= \sup_{\|\alpha\|_0 \leq \widehat{m}, \|\alpha\| \leq 1} \sqrt{|\alpha' X'X\alpha|} \|X(\beta_0 - \widehat{\beta})\| \\
&\leq n\sqrt{\phi(\widehat{m})}\|\beta_0 - \widehat{\beta}\|_{2,n},
\end{aligned}$$

and similarly $\|(X'R)_{\widehat{T}\setminus T}\| \leq n\sqrt{\phi(\widehat{m})}c_s$.

Since $\lambda/c \geq n\|S\|_\infty$, and by Lemma 2, $\|\beta_0 - \widehat{\beta}\|_{2,n} \leq \left(1 + \frac{1}{c}\right)\frac{\lambda\sqrt{s}}{n\kappa_1} + 2c_s$ we have

$$(1 - 1/c)\sqrt{\widehat{m}} \leq 2\sqrt{\phi(\widehat{m})}(1 + 1/c)\sqrt{s}/\kappa_1 + 6\sqrt{\phi(\widehat{m})}\, nc_s/\lambda.$$

The result follows by noting that $(1 - 1/c) = 2/(\bar{c} + 1)$ by definition of $\bar{c}$. □

Proof of Lemma 8. Let $W := \mathbb{E}_n[x_i x_i']$ and $\bar{\alpha}$ be such that $\phi(\lceil \ell k \rceil) = \bar{\alpha}' W \bar{\alpha}$. We can decompose

$$\bar{\alpha} = \sum_{i=1}^{\lceil \ell \rceil} \alpha_i, \text{ with } \sum_{i=1}^{\lceil \ell \rceil} \|\alpha_{iT^c}\|_0 = \|\bar{\alpha}_{T^c}\|_0 \text{ and } \alpha_{iT} = \bar{\alpha}_T / \lceil \ell \rceil,$$

where we can choose α_i's such that $\|\alpha_{iT^c}\|_0 \leq k$ for each $i = 1, ..., \lceil \ell \rceil$, since $\lceil \ell \rceil k \geq \lceil \ell k \rceil$. Since W is positive semi-definite, $\alpha_i' W \alpha_i + \alpha_j' W \alpha_j \geq 2|\alpha_i' W \alpha_j|$ for any pair (i, j). Therefore

$$\begin{aligned}
\phi(\lceil \ell k \rceil) = \bar{\alpha}' W \bar{\alpha} &= \sum_{i=1}^{\lceil \ell \rceil} \sum_{j=1}^{\lceil \ell \rceil} \alpha_i' W \alpha_j \\
\leq \sum_{i=1}^{\lceil \ell \rceil} \sum_{j=1}^{\lceil \ell \rceil} & \frac{\alpha_i' W \alpha_i + \alpha_j' W \alpha_j}{2} = \lceil \ell \rceil \sum_{i=1}^{\lceil \ell \rceil} \alpha_i' W \alpha_i \\
\leq \lceil \ell \rceil \sum_{i=1}^{\lceil \ell \rceil} & \|\alpha_i\|^2 \phi(\|\alpha_{iT^c}\|_0) \leq \lceil \ell \rceil \max_{i=1,\ldots,\lceil \ell \rceil} \phi(\|\alpha_{iT^c}\|_0) \leq \lceil \ell \rceil \phi(k),
\end{aligned}$$

where we used that

$$\sum_{i=1}^{\lceil \ell \rceil} \|\alpha_i\|^2 = \sum_{i=1}^{\lceil \ell \rceil} (\|\alpha_{iT}\|^2 + \|\alpha_{iT^c}\|^2) = \frac{\|\bar{\alpha}_T\|^2}{\lceil \ell \rceil} + \sum_{i=1}^{\lceil \ell \rceil} \|\alpha_{iT^c}\|^2 \leq \|\bar{\alpha}\|^2 = 1.$$

□

Proof of Lemma 9. First note that $P(\max_{1\leq j\leq p} |\widehat{\sigma}_j - 1| \leq 1/4) \to 1$ as n grows under the side condition on n. Let $c_*(m)$ and $c^*(m)$ denote the minimum and maximum m-sparse eigenvalues associated with $\mathbb{E}_n[\tilde{x}_i \tilde{x}_i']$ (unnormalized covariates). It follows that $\phi(m) \leq \max_{1\leq j\leq p} \widehat{\sigma}_j^2 c^*(m+s)$ and $\widetilde{\kappa}(m)^2 \geq \min_{1\leq j\leq p} \widehat{\sigma}_j^2 c_*(m+s)$. Thus, the bound on $\phi(m)$ and $\widetilde{\kappa}(m)^2$ follows from [20]'s proof of Proposition 2 (i) with $\epsilon_1 = 1/3$, $\epsilon_2 = 1/2$, and $\epsilon_3 = \epsilon_4 = 1/16$, which bounds the deviation of $c_*(m+s)$ and $c^*(m+s)$ from their population counterparts. The bound on the restricted eigenvalue κ_1 follows from Lemma 3 (ii) in [2]. Let $M = (s \log(n/e)) \wedge (n/[16 \log p])$ so that as n grows $M/s \to \infty$ under the side condition on s, and we have $M \in \mathcal{M}$ for n sufficiently large since κ_1 is bounded from below and $\phi(M)$ is bounded from above with probability going to one. Thus, the bound on $\widehat{m}$ then follows from Theorem 5 if $\lambda \geq cn\|S\|_\infty$ which occurs with probability at least $1 - \alpha$. □

Proof of Lemma 10. First note that $P(\max_{1\leq j\leq p} |\widehat{\sigma}_j - 1| \leq 1/4) \to 1$ as n grows under the side condition on n. Let $c_*(m)$ and $c^*(m)$ denote the minimum and maximum m-sparse eigenvalues associated with $\mathbb{E}_n[\tilde{x}_i \tilde{x}_i']$ (unnormalized covariates). It follows that $\phi(m) \leq \max_{1\leq j\leq p} \widehat{\sigma}_j^2 c^*(m+s)$ and $\widetilde{\kappa}(m)^2 \geq \min_{1\leq j\leq p} \widehat{\sigma}_j^2 c_*(m+s)$. Thus, the bound on $\phi(m)$ and $\widetilde{\kappa}(m)^2$ follows from [20]'s proof of Proposition 2 (ii) with $\tau_* = 1/2$ and $\tau^* = 2$, which bounds the deviation of $c_*(m+s)$ and $c^*(m+s)$ from their population counterparts. The bound on the restricted eigenvalue κ_1 follows from Lemma 3 (ii) in [2] and the side condition on s. Next let $M = (s \log(n/e)) \wedge ([\epsilon/K_B]\sqrt{n/\log p})$ so that as n grows $M/s \to \infty$, under the side condition on s, and we have $M \in \mathcal{M}$ for n sufficiently large since κ_1 is bounded from below and $\phi(M)$ is bounded from above with probability going to one. Thus, the bound on $\widehat{m}$ then follows from Theorem 5 if $\lambda \geq cn\|S\|_\infty$ which occurs with probability at least $1 - \alpha$. □

References

[1] A. Belloni and V. Chernozhukov (2009). ℓ_1-Penalized Quantile Regression for High Dimensional Sparse Models, arXiv:0904.2931v3 [math.ST].

[2] P. J. Bickel, Y. Ritov and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector, Ann. Statist. Volume 37, Number 4 (2009), 1705-1732.

[3] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp(2006). Aggregation and Sparsity via ℓ_1 Penalized Least Squares, in Proceedings of 19th Annual Conference on Learning Theory (COLT 2006) (G. Lugosi and H. U. Simon, eds.). Lecture Notes in Artificial Intelligence 4005 379-391. Springer, Berlin.

[4] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp (2007). Aggregation for Gaussian regression, The Annals of Statistics, Vol. 35, No. 4, 1674-1697.

[5] F. Bunea, A. Tsybakov, and M. H. Wegkamp (2007). Sparsity oracle inequalities for the Lasso, Electronic Journal of Statistics, Vol. 1, 169-194.

[6] E. Candes and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. Ann. Statist. Volume 35, Number 6, 2313-2351.

[7] D. L. Donoho and J. M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage, Biometrika 1994 81(3):425-455.

[8] R. Dudley (2000). Uniform Cental Limit Theorems, Cambridge Studies in advanced mathematics.

[9] J. Fan and J. Lv. (2008). Sure Independence Screening for Ultra-High Dimensional Feature Space, Journal of the Royal Statistical Society. Series B, vol. 70 (5), pp. 849–911.

[10] V. Koltchinskii (2009). Sparsity in penalized empirical risk minimization, Ann. Inst. H. Poincar Probab. Statist. Volume 45, Number 1, 7-57.

[11] K. Lounici (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators, Electron. J. Statist. Volume 2, 90-102.

[12] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer (2009). Taking Advantage of Sparsity in Multi-Task Learning, arXiv:0903.1468v1 [stat.ML].

[13] N. Meinshausen and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. Annals of Statistics, vol. 37(1), 2246–2270.

[14] P. Rigollet and A. B. Tsybakov (2010). Exponential Screening and optimal rates of sparse estimation, ArXiv arXiv:1003.2654.

[15] M. Rosenbaum and A. B. Tsybakov (2008). Sparse recovery under matrix uncertainty, arXiv:0812.2818v1 [math.ST].

[16] R. Tibshirani (1996). Regression shrinkage and selection via the Lasso. J. Roy. Statist. Soc. Ser. B 58 267-288.

[17] S. A. van de Geer (2008). High-dimensional generalized linear models and the lasso, Annals of Statistics, Vol. 36, No. 2, 614–645.

[18] A. W. van der Vaart and J. A. Wellner (1996). Weak Convergence and Empirical Processes, Springer Series in Statistics.

[19] M. Wainwright (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (Lasso) , IEEE Transactions on Information Theory, 55:2183–2202, May.

[20] C.-H. Zhang and J. Huang (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. Ann. Statist. Volume 36, Number 4, 1567–1594.

[21] P. Zhao and B. Yu (2006). On Model Selection Consistency of Lasso. J. Machine Learning Research, 7 (nov), 2541-2567.